DOCUMENT RESUME

ED 393 936 TM 024 971

| | |
|---|---|
| AUTHOR | Schaeffer, Gary A.; And Others |
| TITLE | The Introduction and Comparability of the Computer Adaptive GRE General Test. GRE Board Professional Report No. 88-08aP. |
| INSTITUTION | Educational Testing Service, Princeton, N.J. |
| SPONS AGENCY | Graduate Record Examinations Board, Princeton, N.J. |
| REPORT NO | ETS-RR-95-20 |
| PUB DATE | Aug 95 |
| NOTE | 56p. |
| PUB TYPE | Reports - Evaluative/Feasibility (142) |
| EDRS PRICE | MF01/PC03 Plus Postage. |
| DESCRIPTORS | *Adaptive Testing; Comparative Analysis; *Computer Assisted Testing; *Equated Scores; Higher Education; *Mathematics Tests; *Test Construction; Test Format; Testing; Test Results; Test Use; Verbal Tests |
| IDENTIFIERS | *Graduate Record Examinations |

ABSTRACT

This report summarizes the results from two studies. The first assessed the comparability of scores derived from linear computer-based (CBT) and computer adaptive (CAT) versions of the three Graduate Record Examinations (GRE) General Test measures. A verbal CAT was taken by 1,507, a quantitative CAT by 1,354, and an analytical CAT by 995 examinees. The verbal and quantitative CATs were found to produce scores that were comparable to their CBT counterparts. However, the analytical CAT produced scores that were judged not to be comparable to the analytical CBT scores. A second study then examined the analytical measure to ascertain the extent of lack of comparability and to obtain statistics that would permit adjustments to restore comparability. Results indicated that the differences in analytical CAT and CBT scores due to the testing paradigm were large enough to require an adjustment in scores. Therefore, in order to enhance the comparability of analytical CAT and CBT scores, the analytical CAT was equated to the analytical CBT. This equating provided new analytical CAT conversions that resulted in comparable analytical CAT and CBT scores. Appendix A explains CBT testing to examinees, and Appendix B presents the CBT program questionnaire. (Contains 10 figures, 20 tables, and 5 references.) (Author/SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED 393 936

GRE
RESEARCH

# The Introduction and Comparability of the Computer Adaptive GRE General Test

Gary A. Schaeffer
Manfred Steffen
Marna L. Golub-Smith
Craig N. Mills
and
Robin Durso

August 1995

GRE Board Professional Report No. 88-08aP
ETS Research Report 95-20

Educational Testing Service, Princeton, New Jersey

BEST COPY AVAILABLE

# The Introduction and Comparability of the Computer Adaptive GRE General Test

Gary A. Schaeffer
Manfred Steffen
Marna L. Golub-Smith
Craig N. Mills
and
Robin Durso

GRE Board Report No. 88-08aP

August 1995

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

Educational Testing Service, Princeton, NJ 08541

**********

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in Graduate Record Examinations Board Reports do not necessarily represent official Graduate Record Examinations Board position or policy.

**********

The Graduate Record Examinations Board and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

EDUCATIONAL TESTING SERVICE, ETS, the ETS logo, GRADUATE RECORD EXAMINATIONS, and GRE are registered trademarks of Educational Testing Service.

Copyright © 1995 by Educational Testing Service. All rights reserved.

**Abstract**

This report summarizes the results from two studies. The first study assessed the comparability of scores derived from linear computer-based (CBT) and computer adaptive (CAT) versions of the three GRE General Test measures. The verbal and quantitative CATs were found to produce scores that were comparable to their CBT counterparts. However, the analytical CAT produced scores that were judged not to be comparable to the analytical CBT scores. As a result, a second study was performed to further examine the analytical measure to ascertain the extent of the lack of comparability and to obtain statistics that would permit adjustments to restore comparability.

Results of the additional study of the analytical measure indicated that the differences in analytical CAT and CBT scores due to the testing paradigm were large enough to require an adjustment in scores. Therefore, in order to enhance the comparability of analytical CAT and CBT scores, the analytical CAT was equated to the analytical CBT. This equating provided new analytical CAT conversions that resulted in comparable analytical CAT and CBT scores.

**Acknowledgments**

A number of individuals provided valuable expertise during this study. The authors thank Tama Braswell and Marion Horta for performing analyses under severe time constraints. Daniel Eignor, Nancy Petersen, and Martha Stocking provided very helpful technical input. Kathleen Carbery served as an excellent GRE Systems contact. Program directors Charlotte Kuh, Susan Vitella, and Jayme Wheeler successfully coordinated the complex implementation of the CAT. James Carlson, Valerie Folk, and Ida Lawrence provided very useful comments on the near-final version of this report. Other ETS staff members, too numerous to mention, provided thoughtful and careful reviews of the results and of earlier drafts of the report. Of course, any shortcomings of the report are the responsibility of the authors.

Table of Contents


Introduction 1

Comparability 2

Methods 2
CAT Development Work 2
CAT Pools 2
Content Specifications 2
CAT Design and Computer Simulations 2
Number of CAT Items 3
Predicted CAT Reliabilities and CSEMs 4
Item Revisits Not Allowed 6
Time Limits 6
Scoring CBTs and CATs 6
Data Collection Design 6
Examinees 6
Test Centers 7
Introduction of CATs 7
CAT in Last Section 7
Testing Tools 8
Score Reporting 8
Description of CAT Samples 9

Analysis of CAT Comparability 10
Parallelism of CBT and CAT Versions 10
Plots of CAT-CBT Difference Scores 13
Baselines for Assessing Magnitude of CAT-CBT Score Differences 16

CAT Timing 18

Subgroup Analyses 19
Subgroup Score Information 19
Subgroup Timing Information 20

Analyses of CAT Algorithm 23

Questionnaire Results 25

Comparability Conclusions 26

**Table of Contents (continued)**


Additional Study of the Analytical Measure 28
Design 28
Description of the Comparability Analysis Sample 29
Comparability Results 29
Discussion 30
Analytical Equating 30
Analytical Equating Methods 30
Impact of Selected Conversions 32

Final Conclusions and Future Considerations 36

References 40

Appendix A: Information for GRE Computer-Based Test (CBT) Examinees 41

Appendix B: Computer-Based Testing Program Questionnaire 43

## Introduction

In June 1988, the Graduate Records Examinations (GRE) Board began consideration of a framework for research and development of a potential new Graduate Record Examination. The Board funded a research and development project to produce a computer adaptive test (CAT) version of the General Test. The project was conducted in two phases because it was recognized that the development of a CAT involves two distinct changes in the presentation of the test. First, the mode of testing is changed. That is, instead of paper and pencil (P&P), a computer is used to present items and record examinee responses. Second, the testing paradigm is changed from a linear test, where all examinees are administered the same set of items, to an adaptive one, where examinees are administered questions that are tailored to their ability. Therefore, the first phase compared a linear P&P test to its linear computer-based test (CBT) counterpart. This comparison addressed effects due to mode of testing. The second phase compared a CAT to a linear CBT. This second comparison addressed testing paradigm effects.

As part of the first phase, a field test was conducted in the fall of 1991 in which a single CBT form was compared to its P&P version. Among the conclusions drawn from this study were (a) examinees were able to navigate through the CBT with little difficulty and their overall reaction to it was favorable and (b) the psychometric characteristics of the linear CBT form were similar to those of its P&P counterpart (Schaeffer, Reese, Steffen, McKinley, & Mills, 1993). Although small numbers of examinees from minority subgroups were included, the study also found no impact on gender and ethnic subgroups as a result of moving from P&P to CBT mode. Equating results supported the use of the same score conversions for the P&P and CBT versions of the test. The scores obtained in the P&P and CBT testing modes were considered to be comparable.

Based on the results of this field test, the GRE Board decided to administer CBTs operationally beginning in October 1992. Two CBT forms were administered, one of which was the field test form and the other a new linear CBT form. Test sections were administered in scrambled orders to enhance test security. Scores were reported to examinees at the test center, as well as by follow-up official score reports. P&P-derived conversions were used for both CBTs, although it needed to be demonstrated that P&P conversions were appropriate for the new CBT form. After several months of data collection, the new CBT was scaled and equated to its P&P counterpart using item response theory (IRT). These resulting conversions were deemed sufficiently similar to the P&P conversions to justify continued use of the P&P conversions for the new CBT form. This CBT equating study, like the field test study, showed that the P&P conversions were essentially the same as the conversions derived directly from the CBT form. Therefore, it has been assumed that additional CBT forms can be introduced and the corresponding P&P conversions used without further study.

The second major phase of this project was to introduce CAT versions of the three GRE measures. Beginning in March 1993, a verbal, quantitative, or analytical CAT was administered in the seventh (final) section of an examinee's CBT session. The primary purpose of this data collection effort was to verify that the scores derived from a CAT measure had similar characteristics to scores derived from a linear CBT (and thus by inference were similar to those for P&P). This comparability of scores is imperative because, for the next

several years, examinees will have the option of taking the GRE General Test in either P&P or CAT mode. However, while these data provided a strong mechanism for detecting differences (or verifying their absence), they were inadequate for making adjustments should any differences be found. And, the differences found for the analytical measure were deemed sufficiently large to require an adjustment. Thus, an additional data collection effort was undertaken to allow the necessary adjustments to be made.

This report is consequently divided into two parts. The first summarizes the results of the comparability analysis and the second provides a description of the equating adjustments for the analytical measure.

## Comparability

### Methods

#### CAT Developmental Work

Much developmental work occurred before CATs were administered in the field. Some basic decisions needed to be made about the structure and functioning of the CATs.

**CAT pools.** A first step was to identify items for inclusion in initial CAT pools. These items previously had been pretested as part of the P&P program, and had been calibrated with the resultant item parameter estimates put on the GRE scale. There were 512, 516, and 660 items in the initial verbal, quantitative, and analytical CAT pools, respectively. Based on the results of the simulation process (see below), the final verbal, quantitative, and analytical CAT pools contained 381, 348, and 512 items, respectively.

**Content specifications.** Detailed content specifications for each CAT measure were generated. These specifications had approximately the same proportions of each item type in the CAT as in the linear CBTs (and P&P versions). To allow for more efficient assessment of ability, the P&P constraint of administering all items of a common type together was removed (one exception was that items with a common stimulus were administered together). This provided for greater measurement precision with a shorter CAT.

**CAT design and computer simulations.** Because it is intended that the P&P and CAT programs will run concurrently, it is necessary that scores derived from both be interchangeable. The design studies for the CATs were undertaken through the use of simulation procedures. The purpose of the simulation studies was to ensure that the two modes would (a) provide scores that were similar; that is, the CAT would on average produce the same means and variances as a linear CBT form, and (b) provide distributions of scores with similar reliabilities and conditional standard errors of measurement (CSEMs).

The algorithm used for selecting items for inclusion in a GRE CAT is governed, in part, by two criteria: optimal information about examinee ability, and consistency of content with what would have been produced by an expert test assembler. Information about the blend of item types contained in a P&P form is incorporated into the selection algorithm in a direct effort to mimic the P&P test assembly process by means of the CAT algorithm. That is,

to help assure that the CAT is measuring the same constructs as a P&P form, item types on the CAT are administered in approximately the same proportions as in a P&P form. As a consequence, the algorithm performs much like an expert test assembler. Also, the concern over test security is incorporated. The CAT algorithm explicitly controls the proportion of examinees to whom an item can be administered. The goal is that no more than 20% of the examinees will see a given item or stimulus. This goal, however, was not achieved; simulation results produced maximum exposure rates of 22-24% across measures. However, the average exposure rate was about 10% for each measure.

The CAT algorithm is an adaptation of a weighted deviations model (Stocking & Swanson, 1992). Basically, each content specification is a rule explicitly incorporated into the model. Ranges of items are specified for each rule, and each rule is assigned a weight that defines its relative importance or reflects its degree of difficulty to achieve. For example, it might be specified that in each analytical CAT the number of items asking the examinee to identify the condition that weakens the presented argument may range from one to three. Any value outside this range is considered a deviation and added to the deviations accumulated across the other rules. The goal is that the weighted sum of the deviations after the last CAT item has been administered should be near zero.

In order to develop a set of weights that resulted in few rule violations and maintained control over exposure rates, simulation studies were undertaken. In these studies the rule weights and exposure rates were systematically manipulated until "acceptable" CATs were produced. Given finite pool sizes, this was often a matter of finding a set of weights that produced acceptable CATs rather than ideal CATs in all instances. Although the qualifications for an acceptable CAT design were varied, the majority of concerns were over violations of major content rules, predicted CSEMs and reliability, and controlled exposure rates. The final decision on acceptability was made by a team of experts from test development, statistical analysis, and program direction at ETS.

**Number of CAT items.** It was decided that each CAT measure would have a fixed number of items because differential test lengths tend to cause a bias in the final ability estimates (Stocking, 1987). Further, differential test length makes it virtually impossible to control the blend of content administered to each examinee. Computer simulations were conducted for varying numbers of items in each CAT measure. The numbers of items selected for the CATs were dependent on several factors, including (a) content specifications, (b) reliability, and (c) CSEMs. The CATs were administered with the following numbers of items:

Verbal CAT-- 30 items
Quantitative CAT-- 28 items
Analytical CAT-- 35 items

The numbers of items in the linear forms are as follows: verbal, 76 items; quantitative, 60 items; analytical, 50 items.



BEST COPY AVAILABLE

**Predicted CAT reliabilities and CSEMs.** One goal of the CAT design was to configure a CAT that would produce scores with characteristics similar to those derived from a particular P&P base form. CAT true score estimates on the base form were scaled to the GRE score scale. Thus, one reason for the notably shorter measure lengths was the goal of matching, not surpassing, the estimate of reliability. Table 1 summarizes the internal consistency reliability (KR-20) of the P&P base form for each measure and the predicted reliability of its CAT counterpart (based on simulated CATs). Plots of conditional standard errors of measurement (CSEMs) are presented in Figures 1a-1c. Although the reliability estimates are quite similar for each measure, it is worth noting that the measurement precision is not identical across the ability continuum. Compared to the P&P base form, the CAT tended to provide better precision at the lower end of the score distribution and similar precision near the middle of the ability continuum. The CAT also tended to provide better precision at the upper end of the score distribution for the quantitative and analytical measures. The relative improvement near the extremes in conjunction with the sparsity of examinees at the extremes accounts for only a slight increase in the overall reliability of the CAT.

Table 1
Base P&P Form and CAT Reliabilities

| | VERBAL | QUANTITATIVE | ANALYTICAL |
|---|---|---|---|
| Base P&P | 0.890 | 0.922 | 0.889 |
| CAT | 0.902 | 0.927 | 0.894 |

Figure 1a
Verbal CAT and P&P CSEMs

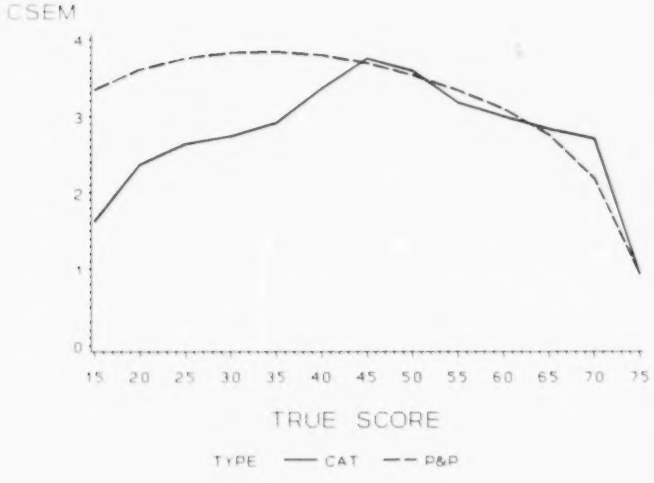


**Figure 1b**
**Quantitative CAT and P&P CSEMs**

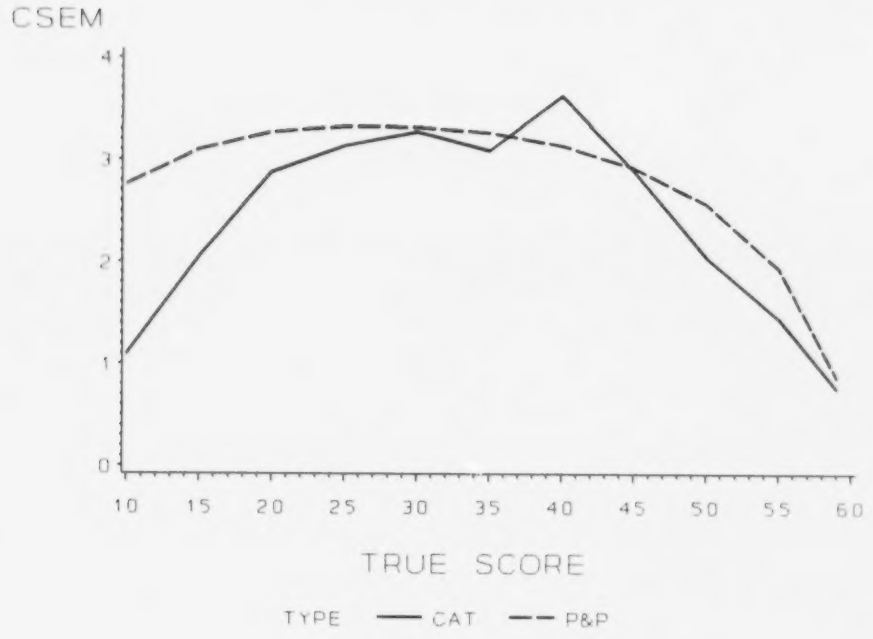
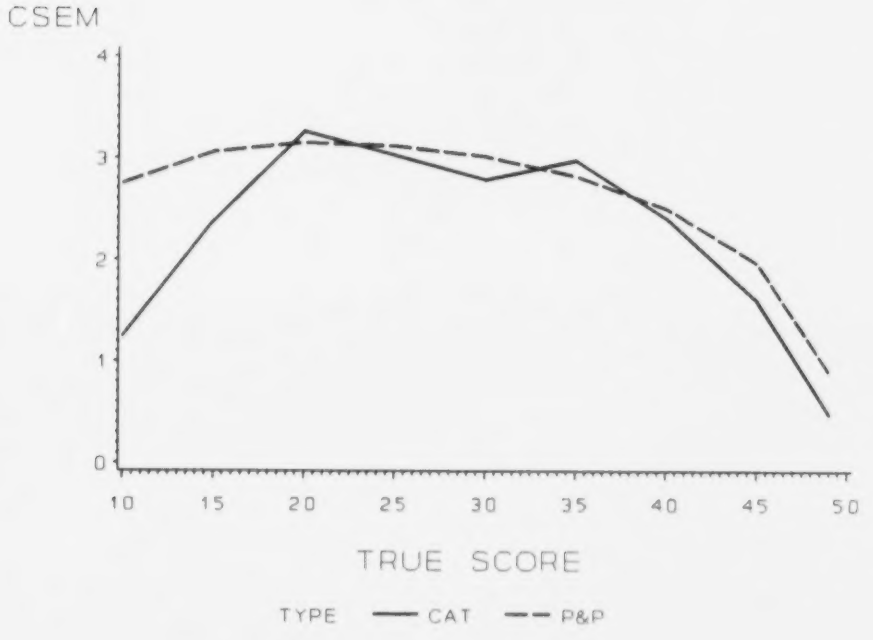



**Figure 1c**
**Analytical CAT and P&P CSEMs**



**Item revisits not allowed.** CAT items were selected for each examinee based on her or his responses to preceding items. For this reason, examinees were not allowed to omit items or revisit preceding items or answers.

**Time limits.** Initial time limits were established for each CAT using the following method. For each measure, a regression model was built that predicted actual CBT field test item times based on examinee ability and item characteristics (Reese, 1993). This model was then applied to CAT simulation data to predict CAT testing times. Distributions of predicted testing times were generated for each CAT measure. Initial CAT time limits used in the present study were selected such that virtually all examinees were predicted to have sufficient time to complete the CAT. These time limits were

Verbal CAT-- 30 minutes
Quantitative CAT-- 45 minutes
Analytical CAT-- 60 minutes

Actual CAT timing data were needed to verify the appropriateness of these time limits, and if the limits had been found to be inappropriate, adjustments would have been made. As reported later in this report, the time limits were found to be appropriate.

## Scoring CBTs and CATs

CBTs and CATs are scored using different methods. Because they are computerized versions of P&P forms, the CBTs were scored number right as is the case with P&P forms. The number-right score was then converted to a scaled score using the corresponding P&P-derived conversion table.

The CATs were scored using an IRT maximum likelihood theta estimation procedure. As an examinee answers each CAT item, the estimate of the examinee's ability is updated based on the examinee's performance on all previous items. At the end of the CAT session, the examinee has a final ability estimate. A table is then used to convert this estimate to an estimate of the number-right true score on the base form, which is then converted to a scaled score. Unlike number-right scoring, this scoring method accounts for the fact that different examinees are administered different items in a CAT, and that some examinees get easier items and some get harder items.

## Data Collection Design

**Examinees.** The subjects of this portion of the study were examinees taking a CBT between March 12, 1993, and September 25, 1993. No special efforts were made to recruit examinees for a CAT administration. Examinees were made aware of the option of taking the GRE on computer from a number of sources, including a supplement to the GRE Bulletin. In addition, beginning in March 1993, a document (see Appendix A) was sent with the registration voucher to all examinees who registered to take a CBT (it also was available at the test sites for walk-in examinees). This document informed examinees that they might get a CAT as the last section of their CBT and described the characteristics of the CAT, including the lack of item-revisit capability. It

also stated that the higher of the CBT and CAT scores would be reported if the examinee met certain test-taking conditions (see Score Reporting section). This was used as an incentive to increase the likelihood of examinees trying their best on the CAT.

**Test centers.** CBT/CAT data were collected from approximately 120 Sylvan test centers, 7 institutions of higher education, and 7 ETS Field Service Offices. Each center had between 4 and 20 work stations, although most centers had 5 or 6. Examinees generally could schedule their test to begin between the hours of 8:00 a.m. and 4:00 p.m.

**Introduction of CATs.** Beginning in March 1993, new scrambled versions of two CBT forms were spiraled at each test center. These were the same two CBT forms that had been used since the CBTs were introduced operationally in October 1992. However, in these scrambled versions different section orders were followed, and either a verbal, quantitative, or analytical CAT appeared in the seventh section. The six scrambled versions of each form were as follows:

| S1 | S2 | S3 | S4 | S5 | S6 |
|---|---|---|---|---|---|
| V1 | V2 | Q1 | Q2 | A1 | A2 |
| A2 | Q2 | V2 | A1 | Q1 | V1 |
| Q1 | A1 | A2 | V2 | V1 | Q2 |
| V2 | V1 | Q2 | Q1 | A2 | A1 |
| A1 | Q1 | V1 | A2 | Q2 | V2 |
| Q2 | A2 | A1 | V1 | V2 | Q1 |
| **V** | **V** | **Q** | **Q** | **A** | **A** |

The measure sections (e.g., V1 and A2) refer to the P&P version sections. The **bold** letters represent the corresponding CAT measure. Thus, one-third of the examinees took each CAT measure in addition to taking three CBT measures.

**CAT in last section.** The design employed in this study had examinees taking a linear CBT for the first six sections and a CAT in the seventh section. The strength of this design was that the same examinees took one measure of the GRE in both linear and adaptive modes, allowing for the comparison of CBT and CAT scores. However, the CAT was always in the last section. This was necessary because examinees were not allowed to revisit items in the CAT but were allowed to do so in the CBT. When examinees went through the test, it was important that they not be asked to switch rules more than once. If the CAT had been in sections 2-6, examinees would have needed to switch rules twice. This would have been undesirable because switching rules during the test could have presented an unnecessary distraction that affected operational scores. The CAT could have been presented in the first section and required only one change of rules; however, it would not have been desirable to start the test with an experimental section.

**Testing Tools**

Once examinees provided sufficient identification at the test center, the center administrator allowed them to begin the test. Examinees used a mouse to navigate on the computer and record responses. Four tutorial sections were presented on the computer to the examinees before the test items were administered. There were tutorials on using the mouse, testing tools, selecting an answer, and scrolling. Examinees could determine how much time they wanted to spend on each tutorial. (Once they left a tutorial they could not return, although tutorial information was available in the Help tool.) The following eight testing tools, each with its own icon, were available to examinees during the CBT portion of the test:

| | |
|---|---|
| Quit: | quit the test |
| Exit: | exit the section |
| Time: | show/hide time remaining in section |
| Review: | go to any item in section/check status of items in section |
| Mark: | mark an item for later review |
| Help: | view previously presented information (i.e., directions, summary of tutorials) |
| Prev: | view screen previously seen |
| Next: | move to next screen |

During the CAT portion of the test, the Review, Mark, and Prev tools were turned off so examinees had to answer each CAT item as it was presented and could not skip items or return to earlier ones. Examinees were informed of this change in tool availability when they began the CAT section.

**Score Reporting**

Rules were devised to encourage examinees to answer as many CAT items as they could. Examinees were told that their CAT score would be reported if it was higher than the linear CBT score and they had either answered all of the CAT items or answered at least 80% of the CAT items before time expired. This decision was based on data indicating that CAT scores from a minimum of 80% of the items provided adequate content representativeness and psychometric characteristics (e.g., reliability and conditional standard errors of measurement), whereas CAT scores based on fewer items generally did not.

Examinees were made aware of these rules in the document distributed with the registration voucher, and by general information screens that appeared on their computer monitors before the CAT began. If one of the two conditions was met, the software compared the CAT score with the CBT score of the like measure and the higher of the two was reported. Otherwise, the CBT score was reported.

At the end of the session, examinees were shown their three scaled scores on their computer monitors. Two of the scores came from the CBT, but for the third score, there was no indication of whether it was from the CBT or CAT. Official score reports were distributed to examinees and designated institutions approximately 12 days after testing. Those sent to examinees listed the number of items scored right, wrong, omit, and not reached for all

CBT scores, information that was not provided for CAT scores. Hence, examinees could determine from their official score reports whether the CBT or CAT score was reported. Official score reports sent to institutions did not indicate at all whether it was a CAT or CBT score.

**Description of CAT Samples**

Most of the analyses were based on a sample of CAT examinees who met certain criteria. Examinees in the analysis sample tested between March 12, 1993, and September 25, 1993. Regular GRE General Test equating sample criteria as well as criteria that indicated that the examinee had a normal testing session and tried to do well on the CAT were used to select the analysis sample. Examinees were selected for the analysis sample if they

1. indicated they were U.S. citizens
2. indicated that they considered English to be their best language
3. marked as a reason for taking the GRE General Test at least one of the following:
   a. admission to graduate school
   b. fellowship application requirement
   c. graduate department requirement
4. had an appropriate irregularity code
5. did not cancel their score
6. had a regularly-timed session
7. had a normal or examinee quit session termination type
8. had a total number of restarts less than or equal to 3
9. had a CAT score computed
10. spent at least one-third of the allotted time on their CAT (as an indication that they were trying to do well)

Of the total 5,221 CBT/CAT examinees who took one of the CBT scrambled versions described earlier, 3,856 (or 74%) met the selection criteria. The majority of examinees not selected into the analysis sample either did not complete the background questionnaire or indicated that they were not U.S. citizens. Of the selected examinees, 1,507 took a verbal CAT, 1,354 a quantitative CAT, and 995 an analytical CAT[1]. The selection criteria did not disproportionately exclude examinees from any gender or ethnic subgroup. Table 2 shows the gender and ethnicity composition of the total selected sample and the sample that took each CAT. Each CAT sample is essentially the same in terms of gender and ethnicity proportions.

[1] The number of examinees taking the analytical CAT is smaller than the numbers taking the other two CATs because some examinees were administered a 29-item analytical CAT instead of the 35-item version as part of a study to determine whether a shorter analytical CAT was viable. The 35-item version was found to be more comparable than the 29-item version.



17 BEST COPY AVAILABLE

Table 2
Gender and Ethnicity Percents

| SAMPLE | N | FEMALE | MALE | AFR.AMER. | ASIAN | HISPANIC | WHITE |
|---|---|---|---|---|---|---|---|
| TOTAL | 3,856 | 57 | 42 | 7 | 3 | 4 | 83 |
| CAT-V | 1,507 | 57 | 43 | 7 | 3 | 4 | 83 |
| CAT-Q | 1,354 | 58 | 42 | 6 | 3 | 4 | 83 |
| CAT-A | 995 | 58 | 42 | 7 | 4 | 3 | 83 |

## Analysis of CAT Comparability[2]

The assessment of CAT comparability involved several analyses. Some analyses addressed how closely the CAT and CBT met the criteria of parallel forms in the classical test theory sense. Other analyses addressed the magnitude of the CAT minus CBT score differences. Baselines were constructed to evaluate these differences.

### Parallelism of CBT and CAT Versions

In classical test theory, two parallel tests have equal observed score means, variances, and correlations with other observed scores. These criteria can be evaluated given that examinees took CBT and CAT versions of the same measure. Table 3 shows the CBT and CAT means and standard deviations for the CAT samples. The first row of CBT scores is for the total sample of all CAT examinees. The remaining rows of scores are for the samples that took each CAT.

Table 3
Score Summary Statistics

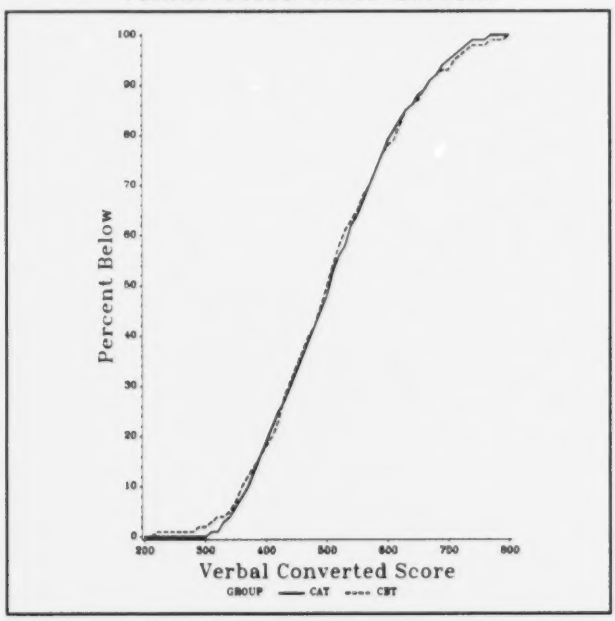
| Sample | N | Mean (and S.D.) of Scaled Scores | | | | | |
|---|---|---|---|---|---|---|---|
| | | VERBAL | | QUANTITATIVE | | ANALYTICAL | |
| | | CBT | CAT | CBT | CAT | CBT | CAT |
| Total | 3,856 | 502 (111) | | 522 (132) | | 543 (130) | |
| CAT-V | 1,507 | 502 (115) | 504 (109) | 522 (132) | | 544 (131) | |
| CAT-Q | 1,354 | 502 (108) | | 522 (131) | 535 (132) | 546 (132) | |
| CAT-A | 995 | 499 (111) | | 522 (132) | | 538 (125) | 555 (135) |

[2]The authors thank Martha Stocking and numerous other ETS reviewers for their input in interpreting the results.

The CAT mean was always higher than the CBT mean. The CAT minus CBT rounded mean differences were 2 for verbal, 12 for quantitative, and 18 for analytical.[3] The standard deviations for the quantitative CAT and CBT were similar. For verbal, the CBT standard deviation was slightly larger than the CAT standard deviation, and for analytical the CAT standard deviation was somewhat larger than the CBT standard deviation.

Figures 2a-2c show score distributions for each measure. The shapes of the CBT and CAT curves for each measure are similar. The CBT curves for the quantitative and analytical measures generally are above the CAT curves, indicating the CAT scores generally were higher than CBT scores.

Figure 2a
Verbal Score Distributions




[3]Note that due to rounding, the mean CAT-CBT differences reported here do not correspond exactly to the differences computed using the means reported in Table 3.

Figure 2b
Quantitative Score Distributions

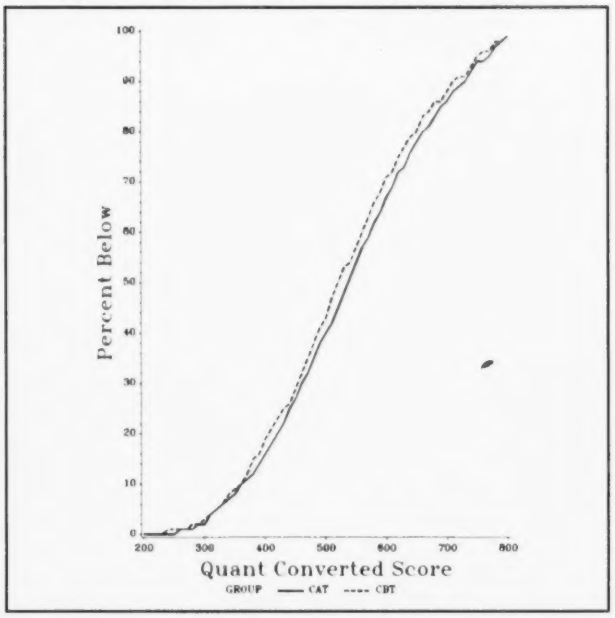



Figure 2c
Analytical Score Distributions

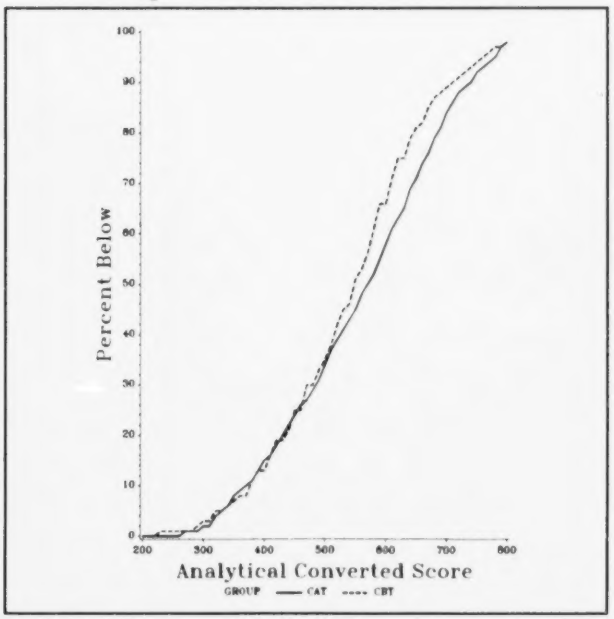


Table 4 shows intercorrelations of CBT and CAT scores with CBT scores for the CAT samples. CBT reliabilities also are presented.

Table 4
CBT and CAT Correlations for the CAT Samples
(decimals omitted; coefficient alpha reliability is underlined)

| | | VERBAL CBT | QUANT CBT | ANALYT CBT |
|---|---|---|---|---|
| VERBAL | CAT | 88 | 52 | 53 |
| | CBT | <u>91</u> | 59 | 60 |
| QUANT | CAT | 51 | 89 | 68 |
| | CBT | 55 | <u>92</u> | 73 |
| ANALYT | CAT | 62 | 70 | 76 |
| | CBT | 62 | 72 | <u>89</u> |

The verbal and quantitative CAT,CBT correlations were only slightly below the CBT reliabilities (.88 versus .91 for verbal, .89 versus .92 for quantitative). The analytical CAT,CBT correlation, however, was somewhat lower than the CBT reliability (.76 versus .89). However, the .89 reliability for the CBT probably is an overestimate of the actual reliability because of the speededness of the test. In addition, the analytical CBT correlations with the verbal and quantitative CBT measures were essentially the same as the analytical CAT correlations with these other two CBT measures. For the verbal and quantitative measures, the CBT,CBT correlations with the other measures were slightly higher than the CAT,CBT correlations.

These data suggest that for the verbal and quantitative measures, the CBT and CAT versions come close to meeting the criteria of parallel forms. The means and standard deviations are similar, the CBT,CAT correlations are only slightly below the respective reliabilities, and the CBT and CAT correlations with other measures are similar (although the CAT correlations are slightly lower than the CBT correlations with other measures). The evidence for parallelism of the CBT and CAT versions of the analytical measure is not as strong. The analytical CAT mean is somewhat higher than the analytical CBT mean, the CAT standard deviation was somewhat larger than the CBT standard deviation, and the analytical CBT,CAT correlation is .13 lower than the analytical CBT reliability. However, the analytical CBT and CAT correlations with other measures are essentially the same.

**<u>Plots of CAT-CBT Difference Scores</u>**

Upon repeated measurement, even with the same instrument, examinees tend to earn different scores. Thus, as expected, examinees taking both the CBT and CAT generally obtain different scores on the two versions. For the CBT and CAT scores to be considered comparable, the differences in CBT and CAT scores generally should be small. CAT minus CBT difference scores were constructed

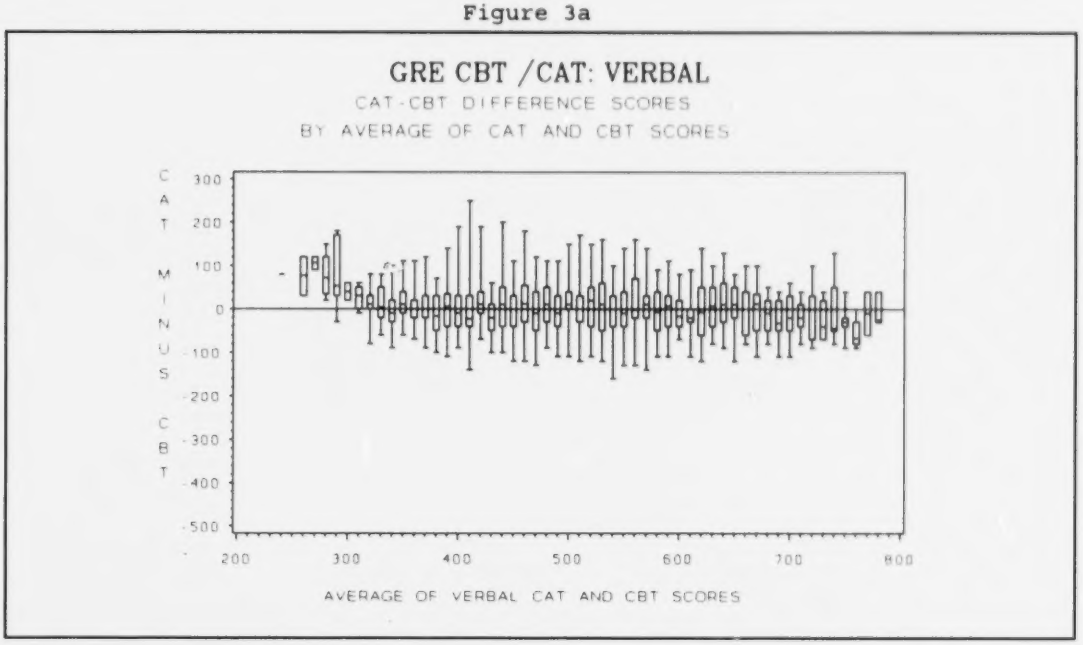


BEST COPY AVAILABLE

for each examinee who took a CAT. Figures 3a-3c show box plots of CAT-CBT score differences plotted against the average of the CBT and CAT scores (rounded to the nearest 10) for the verbal, quantitative, and analytical measures, respectively. The average of the CBT and CAT scores represents examinee ability level. The box plots can be interpreted as follows. The range of scores indicated by the plot represents the range of the difference scores at that ability level. The rectangle represents the interquartile range (25% through 75%) of the difference scores. The median of the distribution of difference scores is represented by a horizontal line within each rectangle.

The box plots illustrate CAT-CBT difference score trends for each measure. A primary concern is the profile of conditional medians. For each measure, the profile is rather flat, particularly where most examinees lie. This suggests that the paradigm impact is similar across the ability continuum. Also, for each measure, the spread of difference scores as represented by the interquartile range is similar across the ability continuum.[4]

Figure 3a




[4]The outlying data point in Figure 3c represents one examinee who had a CAT-CBT difference score of -470 (the rounded average of the CAT and CBT scores was 550). Although this examinee met the analysis sample criteria, the examinee spent only about 24 minutes on the analytical CAT, and, based on CAT item times and item scores, did not appear to employ maximum effort after about the first 20 items.

Figure 3b

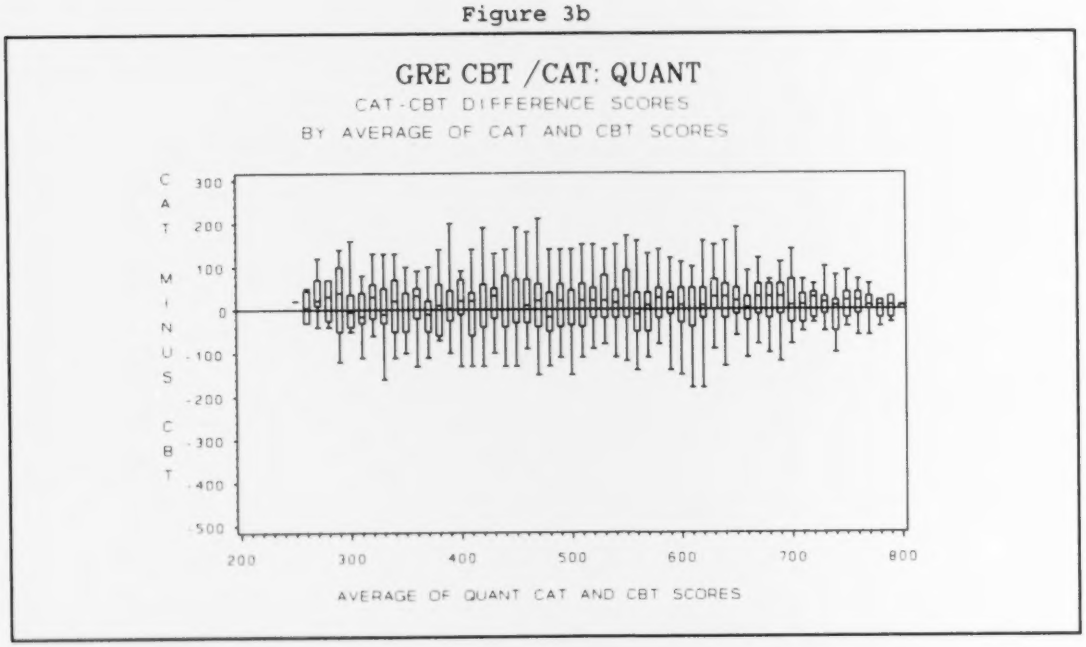
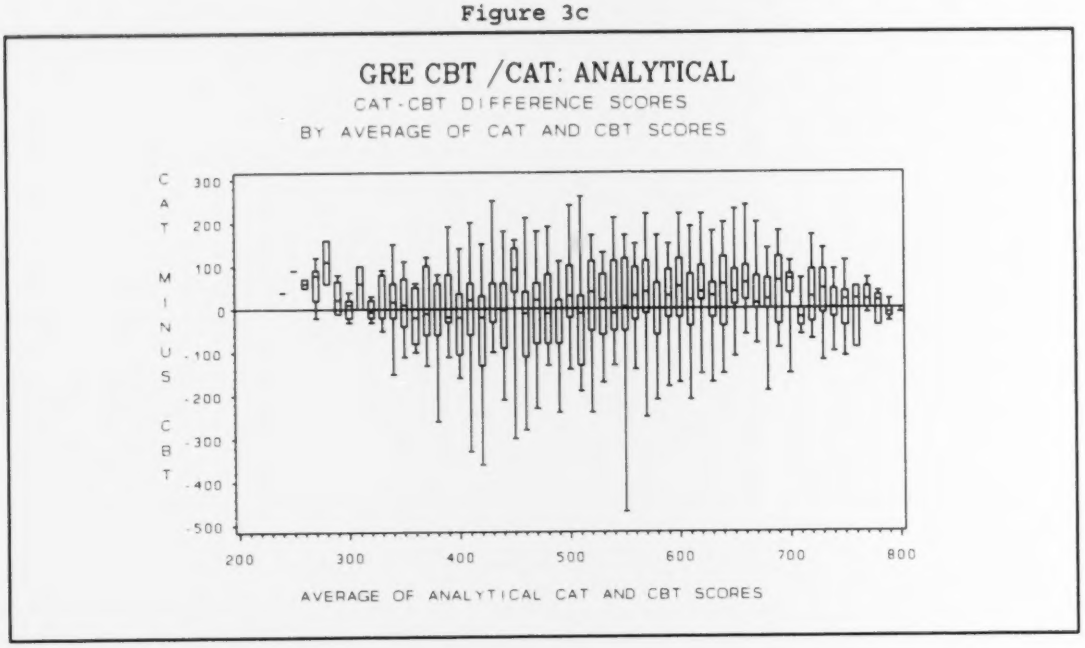



Figure 3c






BEST COPY AVAILABLE

**Baselines for Assessing Magnitude of CAT-CBT Score Differences**

To evaluate the magnitude of differences between the CAT and CBT scores, it was useful to determine the amount of systematic variation that might be expected between two scores derived under similar circumstances. However, no data were available that contained only the repetition of a measure within the same testing session as the source of variation. Two conditions that might bound the circumstances in question were simulation results (the ideal) and natural repeater data (the upper bound).

The magnitude of the differences between CBT and CAT scores was examined in terms of four baselines. The reliability of difference scores for actual data (as opposed to simulated data), however, is extremely low, and therefore caution must be exercised in drawing conclusions based on difference scores alone. In addition, the applicability of repeater baselines is limited because they only somewhat capture the scenario that the examinees followed in the present study. The baselines are

1. Simulated CAT Minus CBT (labeled *SIMUL* in Tables 5-7). Using a population of 8,000 ability parameters with a distribution consistent with a typical December administration, item responses were simulated for both a CAT and a CBT.

2. CBT Field Test (labeled *CBT-P&P*). This baseline includes 1,014 examinees in the fall 1991 CBT field test who took a P&P form at a national administration and then returned several weeks later and took a different form delivered as a CBT.

3. P&P Repeaters from 1992-93 (labeled *93-92*). This baseline includes 1,123 examinees who took different editions of the GRE General P&P test at the December 1992 and February 1993 national administrations.

4. P&P Repeaters from 1981 (labeled *1981*). This baseline includes 498 examinees who took different editions of the GRE General Test in October 1981 and December 1981. This study of GRE repeaters was reported by Kingston and Turner (1984). Some of the data from these repeaters were not available and therefore data from these repeaters were not included in some of the comparisons.

Tables 5-7 provide summary information of the CAT-CBT differences and also of several baselines that were constructed to evaluate the magnitude of those differences. Each row lists a statistic that describes an aspect of the distribution of the difference scores. Each baseline compares the difference of two scores, where, in all cases, the difference score is computed as a gain score, that is, by subtracting the first score from the second score.

For each measure, each CAT-CBT statistic was reasonably close to its baseline counterparts. Some of these findings were noteworthy across measures. For example, the largest mean difference found was for analytical, followed by quantitative and then verbal. Also, the correlation of CAT-CBT scores was smaller for analytical than for verbal and quantitative.

Table 5
VERBAL Baseline Comparisons

| | CAT-CBT | (1) SIMUL | (2) CBT-P&P | (3) 93-92 | (4) 1981 |
|---|---|---|---|---|---|
| Mean Difference | 2 | 1 | 5 | 10 | 23 |
| Difference in S.D. | -6 | -4 | 4 | 0 | * |
| S.D. of Difference Scores | 55 | 47 | 52 | 48 | 63 |
| 5th %ile of Diff. Scores | -90 | -80 | -90 | -80 | * |
| 95th %ile of Diff. Scores | 90 | 80 | 90 | 90 | * |
| Correlation of Scores | .88 | .94 | .90 | .88 | .88 |

* These data were not available for the 1981 baseline.

Table 6
QUANTITATIVE Baseline Comparisons

| | CAT-CBT | (1) SIMUL | (2) CBT-P&P | (3) 93-92 | (4) 1981 |
|---|---|---|---|---|---|
| Mean Difference | 12 | 2 | 1 | 13 | 16 |
| Difference in S.D. | 1 | -3 | 0 | -1 | * |
| S.D. of Difference Scores | 61 | 51 | 55 | 60 | 65 |
| 5th %ile of Diff. Scores | -100 | -80 | -100 | -90 | * |
| 95th %ile of Diff. Scores | 110 | 90 | 90 | 110 | * |
| Correlation of Scores | .89 | .94 | .91 | .88 | .84 |

* These data were not available for the 1981 baseline.

Table 7
ANALYTICAL Baseline Comparisons

| | CAT-CBT | (1) SIMUL | (2) CBT-P&P | (3) 93-92 | (4) 1981 |
|---|---|---|---|---|---|
| Mean Difference | 18 | 4 | 27 | 20 | 17 |
| Difference in S.D. | 9 | -2 | 3 | 7 | * |
| S.D. of Difference Scores | 91 | 59 | 72 | 86 | 78 |
| 5th %ile of Diff. Scores | -140 | -90 | -100 | -120 | * |
| 95th %ile of Diff. Scores | 160 | 100 | 140 | 160 | * |
| Correlation of Scores | .76 | .91 | .83 | .72 | .76 |

* These data were not available for the 1981 baseline.

**CAT Timing**

The initial CAT time limits were set with the intention that almost all examinees would have sufficient time to answer all items. Note, however, that the goal of unspeededness is somewhat in conflict with the comparability goal because the CBT (and P&P tests) are somewhat speeded tests, particularly for the analytical measure. Nonetheless, a goal was for the CAT measures to be less speeded than the CBTs (without much, if any, sacrifice in comparability of scores).

Table 8 presents CAT timing data. Examinees were included who met the analysis sample criteria listed in the Description of CAT Samples section. In addition, examinees who did not answer the minimum number of items needed to compute a CAT score but who used all of the allotted section time were included in this analysis. These selection criteria resulted in slightly greater sample sizes than those listed in Table 3.

The first two rows of Table 8 present the percentages of examinees who answered all and fewer than 80% of the items. A much smaller proportion of CAT analytical (CAT-A) examinees answered all items than did CAT verbal (CAT-V) and CAT quantitative (CAT-Q) examinees. A larger proportion of CAT-A examinees did not answer at least 80% of the total number of items.

Data on timing are presented next. If the test were not speeded, examinees would finish the test early because they could not review. If the test were speeded, examinees would (a) use essentially all the allotted time to complete the test, or (b) fail to complete all items if they paced themselves poorly. The fourth row of the table shows that a large percentage of examinees used all or almost all the allotted time in taking CAT-A. Means and standard deviations of CAT times are presented next, followed by the maximum CAT time allotted. The next-to-last row shows the mean section time divided by the maximum total time allotted. It again appears that CAT-A is

more speeded than the other two CATs.

Additional timing data are presented in the next section on subgroup analyses.

Table 8
CAT Timing Data

| | VERBAL | QUANT | ANALYT |
|---|---|---|---|
| Percentage answering all items | 93 | 88 | 73 |
| Percentage answering <80% of items | 1 | 3 | 6 |
| Total number of items | 30 | 28 | 35 |
| Percentage within 30 sec of max time | 13 | 21 | 42 |
| Mean (and SD) of CAT time in minutes | 24 (4) | 35 (8) | 53 (10) |
| Maximum CAT time allotted in minutes | 30 | 45 | 60 |
| Mean time/maximum time | .80 | .78 | .88 |
| Number of examinees | 1,526 | 1,392 | 1,060 |

## Subgroup Analyses

### Subgroup Score Information

Subgroup sample sizes were sufficient to provide some meaningful descriptive statistics, although larger sample sizes would be required for more thorough analyses. For each CAT sample and subgroup, Table 9 lists the mean and standard deviation of CAT and CBT scores, the number of examinees, and the mean and standard deviation of CAT-CBT rounded difference scores. Almost all subgroups performed better on average on the CAT than on the CBT (the exception was Asian American examinees on the verbal CAT; they performed slightly better on the CBT). CAT-CBT difference scores for female and male examinees were similar for the three measures. Some differences for ethnic subgroups were found. The CAT-CBT difference scores for African American examinees on CAT-V and CAT-Q were positive and much larger than the difference scores for the other subgroups. The CAT-CBT difference score for Asian American examinees on CAT-A was larger than for the other subgroups. Note, however, that this study was not designed to investigate subgroup differences and the numbers of ethnic minority examinees were very small; thus, the generalizability of inferences that can be drawn from these data is limited.

Table 9
Mean and (Standard Deviation) of CAT and CBT Scores by Subgroup*

| Sample | Test | T | F | M | AA | As | H | W |
|---|---|---|---|---|---|---|---|---|
| CAT-V | CAT-V | 504 (109) | 494 (107) | 517 (110) | 428 (101) | 498 (125) | 444 ( 85) | 513 106 |
| | CBT-V | 502 (115) | 492 (110) | 517 (120) | 396 (102) | 507 (119) | 439 (103) | 513 110 |
| | CAT-CBT | 2 ( 55) | 3 ( 56) | 0+ ( 54) | 31 ( 56) | -9 ( 55) | 5 ( 67) | 0- (54) |
| | Number of Examinees | 1,507 | 857 | 649 | 108 | 44 | 60 | 1,249 |
| CAT-Q | CAT-Q | 535 (132) | 501 (118) | 581 (136) | 419 (118) | 605 (129) | 509 (130) | 543 (128) |
| | CBT-Q | 522 (131) | 486 (118) | 572 (132) | 391 (114) | 596 (137) | 490 (126) | 532 (126) |
| | CAT-CBT | 12 ( 61) | 15 ( 62) | 9 ( 60) | 28 ( 59) | 9 ( 56) | 19 ( 59) | 11 ( 61) |
| | Number of Examinees | 1,354 | 782 | 569 | 76 | 35 | 59 | 1,123 |
| CAT-A | CAT-A | 555 (135) | 534 (131) | 585 (133) | 440 (121) | 565 (148) | 509 (148) | 566 (130) |
| | CBT-A | 538 (125) | 516 (121) | 568 (125) | 426 (105) | 521 (168) | 479 (145) | 551 (119) |
| | CAT-CBT | 18 ( 91) | 18 ( 88) | 17 ( 94) | 13 ( 83) | 44 ( 76) | 30 ( 68) | 15 ( 92) |
| | Number of Examinees | 995 | 578 | 416 | 68 | 39 | 33 | 825 |

*T=Total, F=Female, M=Male, AA=African American, As=Asian, H=Hispanic, W=White.

**Subgroup Timing Information**

Table 10 shows, for each gender and ethnic subgroup, the mean and standard deviation of CAT and CBT test times and the percentage of allotted CAT and CBT test times used. Examinees were included who met the analysis sample criteria listed in the Description of CAT Samples section. Also included in this analysis were examinees who did not answer the minimum number of items needed to compute a CAT score but who used all of the allotted section time.

Table 10
Mean, (Standard Deviation), and Mean Percent of Allotted CAT and CBT Test Times In Minutes by Subgroup*

| Sample | Test | T | F | M | AA | As | H | W |
|---|---|---|---|---|---|---|---|---|
| CAT-V | CAT-V | 23.9<br>(4.3)<br>80% | 23.7<br>(4.2)<br>79% | 24.2<br>(4.5)<br>81% | 24.6<br>(4.8)<br>82% | 23.8<br>(3.7)<br>79% | 23.9<br>(4.3)<br>80% | 23.9<br>(4.3)<br>80% |
| | CBT-V | 60.4<br>(6.1)<br>94% | 60.3<br>(5.9)<br>94% | 60.3<br>(6.4)<br>94% | 61.1<br>(5.8)<br>95% | 61.2<br>(4.9)<br>96% | 60.8<br>(4.8)<br>95% | 60.3<br>(6.3)<br>94% |
| | Number of Examinees | 1,526 | 866 | 659 | 111 | 44 | 60 | 1,265 |
| CAT-Q | CAT-Q | 35.0<br>(8.4)<br>78% | 33.9<br>(8.4)<br>75% | 36.4<br>(8.1)<br>81% | 32.5<br>(8.5)<br>72% | 38.8<br>(5.5)<br>86% | 35.2<br>(8.7)<br>78% | 35.1<br>(8.3)<br>78% |
| | CBT-Q | 61.9<br>(5.0)<br>97% | 61.7<br>(5.3)<br>96% | 62.1<br>(4.6)<br>97% | 60.1<br>(7.0)<br>94% | 63.4<br>(1.7)<br>99% | 62.2<br>(3.7)<br>97% | 62.0<br>(4.8)<br>97% |
| | Number of Examinees | 1,392 | 798 | 591 | 78 | 37 | 59 | 1,156 |
| CAT-A | CAT-A | 52.7<br>(9.6)<br>88% | 52.0<br>(9.8)<br>87% | 53.7<br>(9.3)<br>90% | 50.7<br>(10.1)<br>85% | 56.7<br>(4.9)<br>95% | 51.6<br>(10.9)<br>86% | 52.7<br>(9.7)<br>88% |
| | CBT-A | 62.6<br>(4.3)<br>98% | 62.5<br>(4.5)<br>98% | 62.8<br>(3.9)<br>98% | 62.2<br>(4.2)<br>97% | 63.1<br>(4.8)<br>99% | 60.2<br>(10.5)<br>94% | 62.8<br>(3.8)<br>98% |
| | Number of Examinees | 1,060 | 614 | 445 | 69 | 41 | 36 | 881 |

*Allotted times were as follows: CAT-V, 30 minutes; CAT-Q, 45 minutes; CAT-A, 60 minutes. A total of 64 minutes was allotted for each CBT measure.

As can be seen in Table 10, on average all groups spent a larger proportion of allotted time on the CBT than on the CAT, probably because item revisits were allowed on the CBT. A larger mean proportion of the allotted time was spent on the analytical CAT than on the other two CATs, probably because the analytical CAT is more speeded. Female examinees spent on average about 2.5 minutes less on CAT-Q and about 2 minutes less on CAT-A than did male examinees. There were differences among ethnic subgroups on average. On CAT-Q, Asian American examinees spent about 3.5 minutes more and African American examinees about 3 minutes less than Hispanic and White examinees. On CAT-A, Asian American examinees spent about 4-6 minutes more than the other subgroups.

Table 11
Percentage of Examinees Answering Various Numbers of CAT Items by Subgroup*

| CAT | N. Items | T | F | M | AA | As | H | W |
|---|---|---|---|---|---|---|---|---|
| V | <24 | 1 | 1 | 2 | 3 | 0 | 0 | 1 |
| | 24 | 0+ | 0+ | 0 | 0 | 0 | 0 | 0+ |
| | 25 | 1 | 1 | 1 | 1 | 0 | 0 | 1 |
| | 26 | 1 | 1 | 1 | 2 | 0 | 2 | 1 |
| | 27 | 1 | 0+ | 1 | 1 | 2 | 2 | 0+ |
| | 28 | 1 | 0+ | 1 | 5 | 0 | 0 | 0+ |
| | 29 | 1 | 2 | 3 | 5 | 2 | 2 | 2 |
| | 30 | 93 | 94 | 92 | 85 | 95 | 95 | 94 |
| Q | <23 | 3 | 2 | 4 | 3 | 5 | 0 | 3 |
| | 23 | 1 | 1 | 2 | 0 | 0 | 2 | 2 |
| | 24 | 1 | 1 | 2 | 0 | 0 | 0 | 1 |
| | 25 | 1 | 1 | 2 | 0 | 3 | 0 | 1 |
| | 26 | 2 | 1 | 3 | 3 | 11 | 2 | 2 |
| | 27 | 3 | 3 | 4 | 3 | 3 | 5 | 3 |
| | 28 | 88 | 91 | 83 | 92 | 78 | 92 | 88 |
| A | <28 | 6 | 6 | 7 | 1 | 5 | 8 | 6 |
| | 28 | 3 | 3 | 3 | 0 | 5 | 3 | 3 |
| | 29 | 2 | 3 | 1 | 3 | 0 | 0 | 2 |
| | 30 | 3 | 4 | 3 | 7 | 5 | 0 | 3 |
| | 31 | 2 | 2 | 3 | 1 | 7 | 3 | 2 |
| | 32 | 3 | 3 | 3 | 6 | 2 | 8 | 2 |
| | 33 | 2 | 2 | 3 | 3 | 5 | 6 | 2 |
| | 34 | 5 | 4 | 7 | 4 | 10 | 8 | 5 |
| | 35 | 73 | 74 | 71 | 74 | 61 | 64 | 74 |

* T=Total, F=Female, M=Male, AA=African American, As=Asian, H=Hispanic, W=White.

Table 11 shows the percentages of examinees who answered various numbers of items by gender and ethnic subgroups. The examinees in Table 11 are the same examinees that are in Table 10. Male examinees were somewhat less likely to complete CAT-Q than were female examinees. African American examinees were less likely to complete CAT-V than were the other ethnic subgroups. Asian American examinees were less likely to complete CAT-Q than were the other ethnic subgroups. Asian American and Hispanic examinees were less likely to complete CAT-A than were the other two ethnic subgroups. One hypothesis that may explain some of these results is that there is a relationship between item difficulty and time spent on the item. Thus, examinees who are administered

more difficult items may take longer to answer those items and therefore may be less likely to complete the test. There were, however, no marked differences in the percentages of examinees who received scores (all examinees in this table received scores except those in the first row listed for each CAT measure.) Again, note that there were very small numbers of ethnic minority examinees, and this limits the generalizability of comparisons among the subgroups.

## Analyses of the CAT Algorithm

The Methods section describes the process by which the CAT design was established. Final decisions on the design were based on the results from a series of simulations. In addition to assessing the comparability of the linear and CAT versions of each measure, it is important to assess the degree of similarity of the CAT design with actual examinees to the expectations derived from the simulation results. The CAT design strikes a delicate balance among a number of concerns. These include maximum exposure rate (the frequency at which an item is administered); content specifications; overlap constraints (pairs of items or passages that should not be given to the same examinee, e.g., two passages about bicycles); and conditional standard errors of measurement (CSEMs). Any marked deviation from the results obtained from the simulations might indicate that the psychometric characteristics of the CAT with actual examinees differ from expectations and thus require revision. Note that CSEMs cannot be estimated with actual data. Thus, they will not be examined here.

Exposure control parameters in the simulations were adjusted until the maximum exposure rate for any item was as near 0.20 as possible. Given the need to balance exposure control with the other design characteristics, the obtained maximum exposure rates for the simulations were 0.24 for analytical, 0.22 for quantitative, and 0.24 for verbal. Table 12 summarizes the expected and observed usage rates of the CAT pool items. Observed usage rates are summarized for two groups of examinees: those answering all items in the CAT and those receiving a CAT score. That is, data in the "ALL" column are a subset of the corresponding data in the "SCORE" column. Note that the numbers of items in the "ALL" column may be higher or lower than those in the "SCORE" column because the additional examinees in the "SCORE" column could cause an increase or decrease in item exposure rates. Note also that the last two rows compare the "ALL" and "SCORE" results with the simulation results ("SIM").

For examinees answering all items, less than 2% of the items in each pool had observed deviations in exposure rate greater than 5% from expected, and the correlation between expected and observed exposure rates was 0.96 for each CAT. In other words, the most and least frequently used items in the simulations were the most and least frequently used items for actual examinees, respectively. Furthermore, the rates of usage were nearly identical.

Table 12
Simulation and Actual Item Exposure Rates

| ITEM EXPOSURE RATE | VERBAL | | | QUANTITATIVE | | | ANALYTICAL | | |
|---|---|---|---|---|---|---|---|---|---|
| | SIM | ALL | SCORE | SIM | ALL | SCORE | SIM | ALL | SCORE |
| 0.25-0.29 | 0 | 3 | 3 | 0 | 4 | 12 | 0 | 0 | 1 |
| 0.20-0.24 | 37 | 39 | 40 | 51 | 41 | 41 | 20 | 26 | 28 |
| 0.15-0.19 | 64 | 62 | 60 | 50 | 54 | 34 | 74 | 51 | 56 |
| 0.10-0.14 | 35 | 28 | 30 | 20 | 24 | 32 | 59 | 69 | 72 |
| 0.05-0.09 | 53 | 60 | 60 | 52 | 49 | 54 | 96 | 108 | 93 |
| 0.00-0.04 | 93 | 83 | 82 | 95 | 92 | 90 | 152 | 134 | 137 |
| Not used | 68 | 75 | 75 | 62 | 66 | 67 | 48 | 61 | 62 |
| Mean | 0.106 | 0.109 | 0.109 | 0.104 | 0.105 | 0.106 | 0.087 | 0.088 | 0.090 |
| Maximum | 0.240 | 0.271 | 0.276 | 0.218 | 0.261 | 0.282 | 0.244 | 0.249 | 0.260 |
| % diff > .05 | | 0.57 | 0.86 | | 1.52 | 10.00 | | 1.33 | 0.45 |
| Correlation | | 0.965 | 0.964 | | 0.959 | 0.930 | | 0.955 | 0.963 |

Some content specifications were violated for a few simulees in every simulation run. Test development staff reviewed the final simulation runs and found the observed violations to be inconsequential. The results with actual data are virtually identical to the simulation results. For examinees completing the CAT, all violation rates were within 1% of the simulation violation rates, with most rates being identical. The only notable deviations occurred when examinees failed to complete all items and thus were administered fewer items than called for by the CAT design.

Overlap constraints were designed to serve three functions: prohibit the administration of multiple items that essentially test the same logical, mathematical, or linguistic point (structural overlap); prohibit an oversampling of any general field of study (such as business, science, or humanities) so that examinees majoring in any particular field are neither unduly advantaged nor disadvantaged (general subject matter overlap); and prohibit the administration of any two items that happen to mention the same specific ideas, people, or objects (such as depression, Nefertiti, or sailboats) so that the test actually administered to any particular examinee cannot by chance acquire an unintended "theme" (specific subject matter overlap). In no instance did an overlap violation occur in either the simulations or the field.

## Questionnaire Results

At the end of the testing session, each examinees was asked to complete a questionnaire. The questionnaire covered a variety of topics, including prior computer experience, specific reactions to the CBT environment, and CBT and CAT comparisons and preferences. A total of 698 (18%) of the CAT analysis sample examinees completed the questionnaire. Gender proportions were essentially the same in the questionnaire and analysis samples. There were proportionately slightly fewer African American and more White CAT questionnaire respondents than in the analysis sample. The questionnaire respondents had somewhat higher mean scores than the analysis sample. A copy of the questionnaire is in Appendix B.

The questionnaire can be divided into two parts. Questions 1-13 deal with the computer-based testing environment, and questions 14-21 deal with the CATs. On the questionnaire presented in Appendix B, the percentage of all respondents (N=698) selecting different alternatives to each question appears next to the question number for questions 1-13. For questions 14-21, percentages for the verbal, quantitative, and analytical CAT samples are presented separately. For example, 35% of all respondents indicated in question 1 that they used a personal computer some time each week. On question 14, 24% of verbal CAT respondents, 25% of quantitative CAT respondents, and 24% of analytical CAT respondents indicated that they answered all of the questions but felt rushed to do so.

Table B.1 lists the percentages of the total group and of female and male examinees who selected each alternative to each question. Fewer than 23 examinees from any ethnic minority subgroup completed the questionnaire; therefore, results are not presented separately by ethnic subgroup. For questions 1-13, results are presented for the combined CAT samples. For questions 14-21, results are presented separately by CAT sample. For example, 33% of female respondents indicated in question 2 that they owned a IBM/IBM compatible computer. On question 16, 75% of male examinees who were administered a quantitative CAT indicated that they did not care that they were not permitted to review during the last (seventh) section.

As can be seen from responses throughout the questionnaire, the opinions generally were favorable toward the linear CBTs and the CATs. For example, on question 9, 74% of examinees indicated that they thought they would have done as well or better on a CBT as on a P&P test with the same questions. Only about 7% of examinees were very frustrated by not being permitted to revisit or omit items in the CAT (questions 16 and 17). Question 14 indicates that the analytical CAT was perceived as being more speeded than the other CATs. Question 19 indicates that very few CAT examinees thought that many of the questions were too hard or too easy. Responses to question 20 indicate that knowing the minimum number of items required to compute a CAT score affected how examinees worked through the analytical CAT more than it did how they worked through the other CATs. Females and males generally differed only slightly in their responses.



BEST COPY AVAILABLE

## Comparability Conclusions

The purpose of the data collection design for this part of the study was to conduct comparability analyses. Although placing CATs in the last section for all examinees may not have been an optimal design, it was necessitated by a desire for the CAT to function as unobtrusively as possible with regard to an examinee's operational linear CBT performance. Thus, sources of variation such as practice effects are confounded with the effects of adaptive versus linear item administration examined in this study. However, the design allowed the same examinees to take both a linear and an adaptive test, and permitted a direct evaluation of the questions of interest. That is, the data provide a good opportunity to evaluate the CAT algorithm and resulting scores.

Conclusions can be summarized with respect to two questions. First: Is the CAT as delivered in the field consistent with the CAT delivered in simulations? This question addresses whether the construct being measured is the intended one. Second: Are scores obtained consistent across testing paradigms (i.e., linear versus adaptive)? This is a critical question because GRE examinees will have the option of taking the test in either mode (i.e., P&P or computer) and their scores will be compared for high-stakes purposes (e.g., graduate admissions.)

The comparability of the CAT to the CBT was evaluated in terms of several factors. Table 13 lists the factors considered in this study. Each CAT was judged to be at least reasonably comparable to its CBT counterpart in terms of each of these factors, although the analytical CAT measure provided the most mixed results.

Table 13
Comparability Factors

| |
|---|
| Content balance (page 24) |
| Reliability (Table 1) |
| CSEMs (Figure 1a-1c) |
| Scaled score distributions (Fig. 2a-2c) |
| Correlations within measure (Tables 4-7) |
| Correlations across measures (Table 4) |
| Distr. of difference scores (Fig. 3a-3c) |
| Mean difference (Tables 5-7) |
| Difference in S.D. (Tables 5-7) |
| S.D. of difference scores (Tables 5-7) |
| 5th and 95th percentiles (Tables 5-7) |

In addition to evaluating each indicator of comparability separately, in the final analysis all evidence was considered simultaneously. Although there are no formal benchmarks for evaluating multiple indicators simultaneously, a single recommendation is required for each CAT.

The CAT and the linear CBT verbal measures provided strong evidence that it is reasonable to consider scores from both to be comparable. The means differed by only 2 points, which is well within the range observed for the baseline data. The standard deviations differed by 6, which is slightly larger than the differences between standard deviations for the baselines. The correlation between verbal CAT and verbal CBT scores is just slightly lower than the CBT reliability coefficient, 0.88 versus 0.91. The across-measure correlations were lower for the verbal CAT than for the verbal CBTs, but the differences were only 0.07. The verbal CAT appears to introduce some unique variance into the measurement of verbal reasoning. However, there is no evidence that the construct was altered. Furthermore, Figure 2a presents a clear picture that the two score distributions are virtually identical in location and shape.

With the exception of the mean that differs by 12 from the quantitative CBT mean, the quantitative CAT and CBT measures come close to meeting the criteria of parallel forms. The standard deviations differ by 1, the within-measure correlation is just slightly below the CBT reliability coefficient (0.89 versus 0.92), and the across-measure correlations with the CBT verbal and analytical scores are 0.04 and 0.05 below the correlations for the verbal CBT, respectively. The evidence indicates that the CAT and CBT are measures of the same construct. Figure 2b depicts two score distributions that are identical in shape with a slight shift in location. This shift is consistent with baseline data.

The analytical CAT measure provided the most mixed results. The mean difference of 18 is the largest obtained for the three measures, but still within the deviations observed for the baseline comparisons. This CAT produced the largest difference in standard deviations. The across-measure correlations were essentially the same as those for the linear CBT. The within-measure correlation is 0.76, in contrast to a reliability coefficient of 0.89 for the linear CBT. This finding is not particularly surprising, however, given the apparent speededness of the CBT measure. It is quite likely, therefore, that 0.89 is an inflated estimate of the CBT reliability, and that the 0.76, which is similar to the baseline repeater data, may be a better estimate of the reliability of the analytical linear CBT. Evidence such as content similarity and correlational data suggests that the CBT and CAT versions measure the same construct. Conclusions about the similarity of the location and shape of the score distributions are a bit more tenuous. However, Figure 3c shows that both the median and interquartile ranges of CAT-CBT difference scores tend to be rather similar across ability levels. In addition, the differences are not as dramatic as they appear, given the repeater data and the differential manner in which the scoring is affected by speededness.

Although each CAT is an independent measure, several general conclusions are warranted. First, the CATs administered to examinees are consistent with the CAT simulations. The rates of item usage and the proportion of violations



35 BEST COPY AVAILABLE

for each design constraint are virtually identical, and there are no deviations for content constraints of such factors as number of passages to administer or the number of items administered from each of the major item types. Second, examinees seem to have adequate time to consider and answer every item, with the exception of the analytical CAT. Here, however, a conscious decision was made to maintain comparability by retaining some of the speededness present in the linear measure. Third, based on questionnaire data from a limited sample, examinees were comfortable with the CAT environment and administration rules. As expected, a large proportion of examinees given an analytical CAT reported having insufficient time to complete the measure. Fourth, the profile of CAT performance across subgroups is similar to the profile of linear CBT performance, and there is no evidence of consistent negative impact of the CAT for any subgroup. The ethnic subgroup results, however, were based on very small sample sizes.

The overall comparability conclusions were that the verbal and quantitative CATs were adequately comparable to their linear counterparts so that they could be administered operationally without any adjustments. However, the mean difference found between the analytical CAT and the analytical CBT was too large to ignore. Several reasons were proposed for the magnitude of the observed difference. These included actual paradigm differences, within-session practice effects, and differences due to timing. In this design the effects were inseparable. Thus, in order to remove only the systematic sources of variation, an alternative data collection design was required. The following section describes the data collection design and summarizes the results of the adjustments.

## Additional Study of the Analytical Measure

### Design

A design that allows the practice and paradigm effects to be disentangled presents the CAT and linear CBT in counterbalanced order. This also permits an assessment of whether a practice effect is more prominent for a CAT or a linear CBT version of the measure. Also, the performances of both versions are observed in both a practiced and unpracticed condition.

Beginning in mid-November 1993, the three CAT measures were given operationally. The two analytical sections that comprised a single analytical linear CBT measure were also administered. Table 14 summarizes the order in which each of the five sections was administered within each of two scripts. In this table CATA, CATQ, and CATV represent the analytical, quantitative, and verbal CAT measures, respectively. The two sections that constitute the linear analytical measure are denoted by CBTA1 and CBTA2. Examinees were randomly assigned to one of the two scripts. Note that half of the examinees were administered the CAT version of the analytical measure first and the linear version last; the reverse was true for the remaining half of the examinees. To increase motivation throughout the test session, examinees were informed that the higher of their linear and CAT analytical scores would be reported.

Table 14
Section Orders for the Analytical Study

| Script | Section | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| S7 | CATA | CATQ | CATV | $CBTA_1$ | $CBTA_2$ |
| S8 | $CBTA_1$ | $CBTA_2$ | CATQ | CATV | CATA |

This analysis proceeded in two phases. First, using the counterbalanced design, estimates of the magnitude of the paradigm and practice effects were obtained. Second, because the paradigm effect was nontrivial, scores derived from the analytical CAT were equated to those derived from the linear form.

## Description of the Comparability Analysis Sample

During the first two weeks, a total of 1,875 examinees were randomly assigned to take one of the two counterbalanced test scripts that contained both a linear CBT analytical measure and an analytical CAT. Of these, 1,492 (or 80%) met the analysis sample criteria. These examinees had scores computed for both the CBTA and CATA measures. The gender and ethnicity compositions of the groups taking each script were similar to each other and to those reported earlier. The mean scores were somewhat higher for this sample than those previously reported, which was expected given that these examinees tested in November and early December, the time of year when GRE mean scores are traditionally the highest. The percentage of examinees in each subgroup is shown in Table 15.

Table 15
Gender and Ethnicity Percents for the Analytical Study Sample

| SCRIPT | N | FEMALE | MALE | AFR.AMER. | ASIAN | HISPANIC | WHITE |
|---|---|---|---|---|---|---|---|
| S7 | 765 | 54 | 46 | 6 | 2 | 4 | 84 |
| S8 | 727 | 57 | 43 | 6 | 3 | 4 | 85 |

## Comparability Results

Table 16 summarizes the performances of examinees from the two scripts (S7 and S8) on the two analytical measures. Note that means in the $CATA_1$ and $CBTA_2$ cells represent the examinees administered S7, and means for examinees administered S8 are in the $CBTA_1$ and $CATA_2$ cells. The differences in means within the same column quantify the paradigm effect, and the difference in means within the same row quantifies the practice effect. The paradigm effect is very similar across the two columns (11 and 13). Results for the practice effect are also similar (27 and 25).

Table 16
CBTA and CATA Means (and Standard Deviations)
for the Analytical Study Sample

| TEST | ADMINISTRATION ORDER | | |
|---|---|---|---|
| | 1st | 2nd | mean |
| CATA | 585 (130) | 612 (128) | 599 |
| CBTA | 574 (128) | 599 (130) | 587 |
| mean | 580 | 606 | |

The paradigm effect, the difference in the marginal row means, is 12. The practice effect, the difference in the marginal column means, is 26. Two implications of these results are noteworthy. First, the nonzero paradigm effect indicated that data should continue to be collected to allow for adjustment to the CAT conversion table. Second, the practice effect is not ignorable and should either be controlled or adjusted for.

Note that the difference $CATA_2$ - $CBTA_1$ = 38 is much larger than the mean difference of 18 observed in the earlier analyses reported herein. Although we cannot be certain, two plausible explanations for this finding are (a) fatigue washed out some of the practice effect in the earlier study because there were 3 hours of testing time prior to CATA in the earlier study and only 2.25 of testing time prior to CATA in the present study and (b) the effort expended in becoming comfortable with the CAT paradigm in the earlier study may have reduced the practice effect because CATA was the only CAT measure administered (but not so in the present study).

## Discussion

The purpose of this data collection was to determine whether the paradigm effect identified in the earlier comparability analyses was present when practice effects were controlled for. The observed difference of 12 reported score points (although not 18) indicates a need to make an adjustment in the CAT. Had the presence of a paradigm effect not been confirmed, the data collection would have been terminated. However, because a significant paradigm effect was found, the data collection was continued until mid-January.

## Analytical Equating

### Analytical Equating Methods

Throughout the comparability analyses, it was assumed that evidence of a paradigm effect was an indication that the item parameters as estimated in a P&P environment were not adequately predictive of examinee performance when items are selected via a CAT algorithm. Thus, the maximum likelihood estimate of ability ($\hat{\theta}$) is affected. As a result, the corresponding reported score is affected. A direct, but implausible, solution for rectifying this would be to recalibrate all items during CAT administrations. However, a simpler solution

was available that relied on the manner in which the maximum likelihood estimates of ability were converted to the reporting scale. Each CAT was designed to produce unbiased estimates of the number-right scores on a reference form. Reported scores for a CAT were produced by estimating $\hat{\theta}$ from the items selected for administration, transforming this estimate of ability to the number-right scale $\hat{\tau}$ of the reference form, and then applying the scaling table for the reference form. This can be represented by

$$\{(a_i, b_i, c_i)\} \rightarrow_1 \hat{\theta} \rightarrow_2 \hat{\tau}_{ref} \rightarrow_3 SS.$$

In this model, transformations 1 and 2 are mathematically defined and not really available for adjustment. However, the $\hat{\tau}$ to SS transformation can be adjusted. The purpose of this adjustment is not to correct the transformation from the number-right to the reported scale for the reference form. The presence of a paradigm effect is evidence that the $\hat{\tau}$ derived from the CAT is, in a sense, biased. Thus, the purpose of the adjustment is to find an alterative estimate ($\hat{\tau}_{alt}$) that results in a SS with no paradigm effect present. Alternatively, the CAT can be viewed as a form that produces a pseudo raw score ($\hat{\tau}_{ref}$) that has yet to have a scaling transformation defined. Because examinees were administered S7 or S8 at random, the data for a randomly equivalent groups design was available. Differences between the CATA and CBTA scores were not uniform across the ability scale. Consequently, an equipercentile equating of the $\hat{\tau}_{ref}$'s from the CAT to the observed number-right score on the CBT form was used to eliminate the paradigm effect.

Table 17 presents the CBTA and CATA means and standard deviations by administration order for the equating sample. Examinees in the equating sample tested between mid-November 1993 and mid-January 1994. The equating sample sizes for the two scripts were 3,543 and 3,600 for scripts S7 and S8, respectively. Once again, the means are presented in this fashion to help illustrate the magnitude of the paradigm and practice effects. The overall difference ($CATA_2$ - $CBTA_1$) of 39 is similar to that for the initial data. However, here the paradigm effect taking into account both the practiced and unpracticed data is 16 and the practice effect is 23. The paradigm effect taking into account only the data not affected by practice is 20 ($CATA_1$ - $CBTA_1$).

Table 17
CATA and CBTA Means (and Standard Deviations)
for the Equating Sample

| TEST | ADMINISTRATION ORDER | | |
|---|---|---|---|
| | 1st | 2nd | average |
| CATA | 593 (128) | 612 (127) | 603 |
| CBTA | 573 (126) | 601 (125) | 587 |
| average | 583 | 606 | |



BEST COPY AVAILABLE

The obtained sample sizes (≈3,600) were only of moderate size for performing equipercentile equatings. Thus, the frequency distributions were smoothed using a log-linear smoothing technique holding from two to five moments fixed. From the four smoothings for each distribution, the smoothing that was judged to best represent the original frequency distribution was selected for use during equating. Equatings were performed with the unpracticed, practiced, and pooled data. However, it was believed a priori that the equating based on the unpracticed data would most cleanly eliminate the paradigm effect in question. The other equatings were run to confirm that the results so derived would not be markedly different. Nothing in the results contradicted the a priori position. Consequently, the conversions based on the unpracticed data were selected for use.

**Impact of Selected Conversions**

Figure 4 displays the original CATA and the equated CATA conversion functions. The equated CATA conversion produces lower scores throughout the score range.

Table 18 shows CBTA, equated CATA, and original CATA summary statistics for the unpracticed data from the equating sample. The CBTA column represents examinees who were administered CBTA first, and the two CATA columns represent examinees who were administered CATA first. As expected, the equated CATA statistics were more similar to the CBTA statistics than were the original CATA statistics. The correlation between the equated CATA and original CATA scores was 0.996.

**Figure 4**
**Equated CATA and Original CATA Conversion Functions**

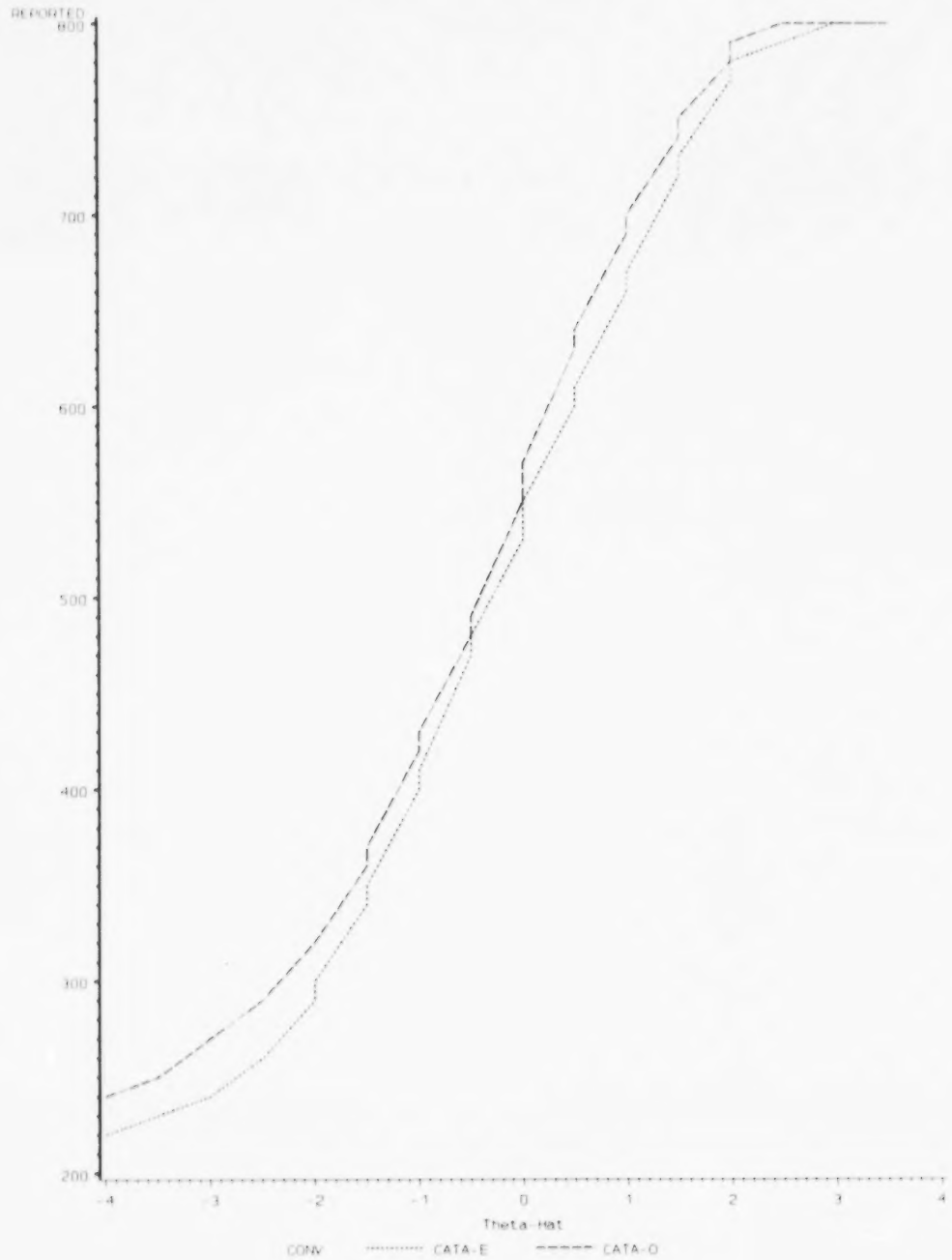


Table 18
Summary Statistics for Unpracticed Data for the Equating Sample*

| | CBTA | CATA-E | CATA-O |
|---|---|---|---|
| N. EXAMINEES | 3,600 | 3,543 | 3,543 |
| MEAN | 573 | 573 | 593 |
| STD. DEVIATION | 126 | 126 | 128 |
| SKEWNESS | -0.33 | -0.33 | -0.37 |
| KURTOSIS | -0.38 | -0.38 | -0.62 |
| 10TH PERCENTILE | 400 | 400 | 410 |
| 25TH PERCENTILE | 490 | 490 | 500 |
| 50TH PERCENTILE | 590 | 580 | 610 |
| 75TH PERCENTILE | 660 | 660 | 690 |
| 90TH PERCENTILE | 740 | 740 | 760 |

*CATA-E refers to the equated analytical CAT score and CATA-O refers to the original analytical CAT score.

Table 19 shows the means and standard deviations of equated CATA scores minus original CATA scores for the equating sample from the two test scripts and for gender and ethnic subgroups. The effect of the equating in reducing the CATA scores was similar for each of the subgroups.

Table 19
Equated CATA Minus Original CATA Difference Score Statistics

| | TOTAL | FEMALE | MALE | AFR.AMER. | ASIAN | HISP. | WHITE |
|---|---|---|---|---|---|---|---|
| MEAN | -20 | -20 | -20 | -19 | -20 | -20 | -20 |
| STD.DEV. | 9 | 8 | 9 | 8 | 9 | 8 | 9 |
| NUMBER | 7,143 | 3,783 | 3,348 | 427 | 232 | 303 | 5,942 |

Table 20 shows the percent distribution of examinees with specified equated CATA minus original CATA difference scores conditioned on grouped analytical ability, where analytical ability is defined as their score from the analytical measure taken first (either CBTA or equated CATA). All changes are within -40 to 0 reported scale score points; 98% of the changes are within -30 to -10 scaled score points.

Table 20
Percent Distribution of Equated CATA Scores Minus Original CATA Scores

| | DIFFERENCE (CATA-E - CATA-O) | | | | | |
|---|---|---|---|---|---|---|
| ABILITY* | -40 | -30 | -20 | -10 | 0 | FREQ |
| 790-800 | 0 | 9 | 17 | 53 | 21 | 378 |
| 770-780 | 0 | 9 | 11 | 78 | 2 | 201 |
| 750-760 | 0 | 15 | 41 | 43 | 0 | 256 |
| 730-740 | 1 | 27 | 56 | 15 | 1 | 400 |
| 710-720 | 0 | 28 | 68 | 5 | 0 | 280 |
| 690-700 | 1 | 47 | 48 | 4 | 0 | 298 |
| 670-690 | 0 | 68 | 27 | 4 | 0 | 454 |
| 650-660 | 1 | 75 | 16 | 7 | 0 | 506 |
| 630-640 | 1 | 72 | 17 | 11 | 0 | 489 |
| 610-620 | 0 | 73 | 17 | 10 | 0 | 366 |
| 590-600 | 0 | 58 | 25 | 16 | 0 | 492 |
| 570-580 | 0 | 45 | 39 | 16 | 0 | 350 |
| 550-560 | 0 | 20 | 57 | 24 | 0 | 459 |
| 530-540 | 0 | 8 | 47 | 45 | 0 | 311 |
| 510-520 | 0 | 10 | 27 | 63 | 0 | 390 |
| 490-500 | 0 | 4 | 10 | 86 | 0 | 228 |
| 470-480 | 0 | 5 | 15 | 80 | 1 | 171 |
| 450-460 | 0 | 7 | 22 | 71 | 0 | 246 |
| 430-440 | 0 | 4 | 23 | 73 | 0 | 133 |
| 410-420 | 0 | 3 | 29 | 68 | 0 | 120 |
| 390-400 | 0 | 4 | 24 | 75 | 0 | 116 |
| 370-380 | 0 | 20 | 28 | 52 | 0 | 130 |
| 350-360 | 0 | 10 | 56 | 35 | 0 | 63 |
| 330-340 | 0 | 17 | 68 | 15 | 0 | 60 |
| 310-320 | 0 | 36 | 53 | 12 | 0 | 59 |
| 290-300 | 0 | 55 | 35 | 10 | 0 | 51 |
| 270-280 | 0 | 60 | 20 | 18 | 2 | 50 |
| 250-260 | 0 | 65 | 29 | 6 | 0 | 51 |
| 200-240 | 0 | 43 | 40 | 17 | 0 | 35 |
| TOTAL | 21 | 2,533 | 2,274 | 2,225 | 90 | 7,143 |

*Analytical ability as defined by the unpracticed analytical score, either CBTA or equated CATA.

Finally, another outcome of this study was the confirmation of the presence of practice effects. Results from the counterbalanced design indicated a rather large practice effect for the analytical measure. This has implications for the future when pretest sections are administered with the operational CATs. The Program is considering various administrative options for reducing or eliminating practice effects from operational and pretest scores.

**Final Conclusions and Future Considerations**

The verbal and quantitative CAT score distributions were found to be sufficiently similar to the respective CBT score distributions that no adjustment was necessary for these CATs to be considered comparable to their CBT counterparts.

Scores on the analytical CAT, however, were sufficiently higher on average than analytical CBT scores to require an equating adjustment. An equating study conducted to derive new analytical CAT conversions resulted in comparable equated CAT scores and CBT scores (as required by the equating), and no differential negative impact was found for subgroups. These new conversions should apply to future analytical CAT pools where the item parameters also will be obtained from P&P administrations.

The completion of the comparability study is a major accomplishment for the GRE Program; however, there are still many issues to be addressed regarding the ongoing operation of a large-scale adaptive testing program. In this section, we list some issues that lie ahead. The following are briefly discussed:

- What is the optimal configuration of pools?
- How can the quality of a pool be monitored and maintained over time?
- What is needed to assure equivalence of computer and paper testing in international settings?
- What opportunities and problems do computer adaptive tests create with regard to testing individuals with disabilities?
- How can pretesting be accomplished in a computer adaptive testing program? Are current techniques for evaluating pretest results adequate?
- Will adaptive testing result in differences in traditional patterns of differences among subgroups?
- What is the effect of administrative procedures such as the lack of review in adaptive tests?

*What is the optimal configuration of pools?*

In traditional testing programs, one set of questions is administered to large numbers of persons on a single day. Thus, item exposure is limited to a short period of time. In adaptive testing, however, the period of time in which items are exposed is increased, although the rate of exposure may be lessened. In the short term, this appears to enhance test security. There will be less incentive to memorize a given adaptive test item because there is no guarantee that another test taker will receive the same (or mostly the same) items.

In the longer term, however, even a low exposure rate can mean a high exposure volume. If, for example test questions are exposed to 10% of a testing programs volume, 100,000 examinees will have seen an item after a million have been tested. If CAT pools are to be in operation for long periods of time, this level of exposure would become commonplace (in GRE, it would take only about three years to reach a million examinees). A question to be addressed, then, is what is the most effective way to reduce item exposure. Should items continue to be added to a single pool, thus lowering the exposure rate within the pool, or should multiple pools with constant exposure rates within each pool be developed? If multiple pools are developed, how many are needed? Can items be used in more than one pool?

*How can the quality of a pool be monitored and maintained over time?*

Little is known about the extent to which items retain their characteristics upon repeated administrations. It is conceivable that questions will change in quality at different rates. Some questions may, for example, be particularly memorable and become known quickly. Others may be selected for administration at a high rate and need to be removed from the pool to avoid overexposing them.

Removal of items that are selected most often may pose a problem for pool maintenance because the selected items are likely to be those of highest quality. If pretesting cannot yield sufficient volumes of highest quality items, the psychometric quality of the pool will degrade over time. That is, the number of items required may need to increase.

To date, little is known about how to monitor item quality over time in adaptive tests. Item parameters were developed on a sample with a wide range of abilities, but the items will be administered to individuals with a more narrow range of ability. As a result, monitoring the stability of parameters over time may be difficult. However, some mechanism is required that will allow programs to monitor exposure rates and item performance to determine when items need to be replaced.

*What is needed to assure equivalence of computer and paper testing in international settings?*

Although the research conducted to date has demonstrated that adaptive and traditional tests can be comparable, the expansion of computer adaptive testing throughout the world raises new questions. Our research indicated that people with little or no computer familiarity can learn the testing system and use it effectively in a short period of time. However, in the United States people are

quite familiar with technology (e.g., ATMs) It is not clear that these results will necessarily hold in countries and regions where technology is not as widespread.

*What opportunities and problems do computer adaptive tests create with regard to testing individuals with disabilities?*

The potential of the computer to provide alternatives to traditional test modifications is apparent. Many types of "alternative" input devices are already available. Multimedia offers the potential for recording tests or providing standardized American Sign Language presentations. Screen displays can be altered easily (e.g., changing color or magnifying type). As these modifications are incorporated into test delivery systems, there may be debate about whether or not they constitute a modification. If, for example, most commercial software packages allow the user to modify colors, is changing the color for a testing application a modification that should be identified on the score report? If not, should it be generally available to all test takers?

Other questions that are likely to arise include how to administer adaptive tests in Braille format, whether using a speech synthesizer alters the construct being measured by a reading comprehension test, and so forth. Traditional definitions of "standard" administrations may be called into question.

*How can pretesting be accomplished in a computer adaptive testing program? Are current techniques for evaluating pretest results adequate?*

In traditional tests, pretesting is usually accomplished either through an unidentified, separately timed section or through the embedding of pretest questions within the operational test. Both methods are also available in adaptive testing; however, it is not clear whether one should be preferred over another. With embedded pretests there is a risk of tainting operational performance if a flawed pretest item is administered. Encapsulated sections of pretest items do not run this risk, but are quite difficult to manage in a modular environment. Equivalence of item parameters derived from pretesting in a traditional setting and from adaptive settings must be established. New methods of evaluating pretest data may also be required.

*Will adaptive testing result in differences in traditional patterns of differences among subgroups?*

Although the results of the comparability study demonstrated that we can achieve comparability of traditional and adaptive tests, data on subgroup performance were limited. There are three possible outcomes of adaptive tests with regard to subgroup performance. First, there may be no change in traditional relationships among groups. Second, score differences may increase. This concern has been widely expressed given differential access to computers. Third, score differences may decline. It is possible to hypothesize that traditional tests that are inappropriately difficult for some people may be sufficiently frustrating to them that performance is depressed. Targeting tests to performance may remove that source of variance and result in higher scores. Clearly, performance of subgroups should receive special scrutiny for adaptive tests.

*What is the effect of administrative procedures such as the lack of review in adaptive tests?*

Although it is possible to administer adaptive tests and allow item review, the GRE adaptive test does not allow review. This administrative decision was made to (1) allow administration of tests that were as short as possible and (2) discourage test takers from deliberately missing questions to obtain an easy test and then revise their answers in the hope of obtaining a very high score on a very easy test. However, prohibiting review is of concern to individuals who posit that test takers will continue to consider test questions after they have answered them with the occasional result that they remember something that allows them to answer correctly an item they previously missed. It is unclear whether review is important to the validity of the test. The results of this investigation suggest that it is not because the scores were comparable, but they are not conclusive. Additional research is necessary to determine the importance of review and, if it is important, to determine ways of allowing it without the potential of degrading the psychometric quality of the test.



BEST COPY AVAILABLE

## References


Kingston, N.M., & Turner, N. (1984). *Analysis of score change patterns of examinees repeating the Graduate Record Examinations General Test* (Research Report No. RR-84-22). Princeton, NJ: Educational Testing Service.

Reese, C.M. (1993). *Establishing time limits for the GRE computer adaptive test*. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Schaeffer, G.A., Reese, C.M., Steffen, M., McKinley, R.L., & Mills, C.N. (1993). *Field test of a computer-based GRE General Test* (Research Report No. RR-93-07). Princeton, NJ: Educational Testing Service.

Stocking, M. S. (1987). Two simulated feasibility studies in computerized adaptive testing. *Applied Psychology: An International Review, 36*, 263-277.

Stocking, M., & Swanson, L. (1992). *A method for severely constrained item selection in adaptive testing* (Research Report No. RR-92-37). Princeton, NJ: Educational Testing Service.

**Appendix A**

## Information for GRE® Computer-Based Test (CBT) Examinees

Beginning sometime in March 1993 and continuing through at least September 1993, GRE CBT examinees will have the opportunity to participate in an evaluation of a new kind of test called a Computer Adaptive Test (or "CAT"). The main purpose of this evaluation is to try out the CAT in actual CBT centers before it becomes part of the regular GRE computerized testing program. For this evaluation, CBT scores will be derived from the first six sections, and the seventh section will contain either a Verbal, Quantitative, or Analytical CAT, which does not contribute toward your CBT score. However, examinees participating in the adaptive test evaluation will have an opportunity to improve one of their GRE test scores. If the CAT score you earned in Section 7 is higher than your score in the corresponding CBT section, your CAT score will become your official score and will be reported. The particular CAT section you are given will be determined randomly.

*What is computer-adaptive testing?*

Traditionally, examinees who are given the same test form are given the same questions. This occurs in both the paper-and-pencil and CBT formats. However, the easy questions are too easy for some examinees, and the hard questions are too hard for others. In a CAT, everyone starts with a question that is randomly selected from a group of approximately middle difficulty questions. If you answer the first question correctly, the next question given to you will be more difficult, but if your answer is incorrect, the next question will be easier. Throughout the test, questions are selected for you based on your performance on previous questions. The difficulty levels of the questions are known because the questions have been administered previously to GRE examinees. Because you are given only questions that are at an appropriate level of difficulty for you, the CAT consists of fewer questions than the CBT or paper-and-pencil test.

*How is the CAT scored?*

In a CBT or paper-and-pencil GRE General Test, each examinee's score is based on the number of questions answered correctly. In a CAT, where some examinees are given easier questions than other examinees, it would not be appropriate to base each examinee's score solely on the number of questions answered correctly. Consequently, correctly answering difficult questions counts more than correctly answering easy questions. That is, the examinee who correctly answers difficult questions gets a higher CAT score than the examinee who correctly answers the same number of easier questions. However, if you have been given the most difficult questions and answer some of them incorrectly, you can still get a high score.

*How do I proceed through the CAT?*

In the CAT you must answer every question in the order in which it is presented to you. You cannot omit questions, and you cannot return to previous questions. You will NOT be able to use the Previous, Review, and Mark testing tools during the CAT. That is because the questions given to you are based in part on your answers to earlier questions. The questions you are given are being selected for you as you take the test. You can, however, change an answer before you proceed to the next question.

*Can I get "stuck" with the wrong questions?*

If your answer to a question is due to a careless error or a lucky guess, your answers to the following questions will direct you back toward questions at the appropriate level of difficulty for you. The adaptive nature of the CAT allows the test to correct itself, because your answers to all previous questions determine your subsequent questions.

*What about different types of questions?*

In the CAT, not only does every examinee have the same opportunity to be given the hard questions, but every examinee will get questions that are very similar in the mix of content being measured and the types of questions being used. For instance, in the Quantitative CAT, the computer selects about the same number of arithmetic, algebra, and geometry questions for each examinee. Also each question type in a CAT is not necessarily grouped with others of that type as they are in the CBT and the paper-and-pencil tests. For example, an examinee taking the Verbal CAT may be given an analogy question followed by a sentence completion question and then another analogy question.

*What is the best test-taking strategy for a CAT?*

The best strategy is simply to answer each question to the best of your ability. Even though a correct answer will generally be followed by a more difficult question, it is to your advantage to try to answer each question correctly, since difficult questions count more toward getting a higher score.

*Is a CAT score different from a score earned on the paper-and-pencil General Test or the CBT?*

It is anticipated that CAT scores will be interchangeable with scores earned on both the CBT and the paper-and-pencil tests. That is, examinees, on average, would be expected to get very similar scores on the paper-and-pencil test, CBT, and CAT. Also, mode of testing (i.e., paper-and-pencil, CBT, or CAT) will not be indicated on score reports sent to designated institutions. In this evaluation, if your CAT score is higher than your CBT score and the CAT score is, therefore, reported, your examinee score report will not indicate the number of questions you answered correctly or incorrectly on the CAT.

*How long is the CAT?*

One of the purposes of this evaluation is to determine whether the time limits currently established for each measure are appropriate. Depending on which CAT section you receive, you will be given the following numbers of questions and time limits:

| | |
|---|---|
| Verbal CAT: | 30 questions, 30 minutes |
| Quantitative CAT: | 28 questions, 45 minutes |
| Analytical CAT: | 35 questions, 60 minutes |

*What if I still have questions about the CAT?*

At your CBT session, you will be given complete instructions for taking the CAT section right before it is administered. The directions will be presented on the computer and will precede Section 7. You will also be given debriefing material after the testing session.

Appendix B

## COMPUTER-BASED TESTING PROGRAM QUESTIONNAIRE

**Please circle the appropriate response to each question unless noted otherwise.**

1. How often do you use a personal computer?

3 (1) Never before taking the GRE/CBT (Skip to Question 7.)
19 (2) Rarely
35 (3) Some time each week
43 (4) Almost daily

2. Do you own a personal computer?

38 (1) Yes, IBM/IBM Compatible
17 (2) Yes, Mac/Apple
4 (3) Yes, Other
38 (4) No

3. If you answered No to Question 2, do you have a personal computer available for your use?

33 (1) Yes
8 (2) No

4. How would you describe your ability to type using a computer keyboard?

1 (1) No ability
5 (2) Poor
26 (3) Fair
43 (4) Good
23 (5) Excellent

5. During the past year, how often have you used a word processing package to write a report, term paper, letter, etc.?

8 (1) Never
33 (2) From time to time during the year
41 (3) At least once a week
15 (4) Daily

6. How often have you used a mouse on a personal computer?

14 (1) Never before taking this test
30 (2) A few times
28 (3) At least once a week
25 (4) Daily

7. In the CBT, sometimes all of the information cannot be presented on a single screen. When there was information that required scrolling, how apparent was the need to scroll?

58 (1) Very apparent
36 (2) Somewhat apparent
1 (3) Not apparent at all
5 (4) Only apparent after reading the question

**The following two questions ask you to compare your computer-administered test experience to a paper-and-pencil test experience.**

8. How would you compare the computer test-taking experience with taking a paper-and-pencil test?

55 (1) Better than a paper-and-pencil test
25 (2) About the same as a paper-and-pencil test
19 (3) Worse than a paper-and-pencil test

9. How do you think you would have done on a paper-and-pencil test with the same questions?

13 (1) Not as well on the paper-and-pencil test
61 (2) About the same
24 (3) Better on the paper-and-pencil test

**The following questions deal with the test center environment.**

10. How knowledgeable was the test center staff about the CBT administration?

73 (1) Very knowledgeable
16 (2) Somewhat knowledgeable
1 (3) Not knowledgeable
10 (4) I did not ask any questions.

11. Were there any distractions or inconveniences during the testing session? Select as many as apply.

63 (1) No distractions or inconveniences
1 (2) Noisy testing room
1 (3) Inadequate lighting
9 (4) Noise made by other examinees was distracting.
10 (5) Noise made by center staff helping other examinees was distracting.
15 (6) Noise outside the testing room was distracting.
4 (7) The table space was inadequate to do scratch work.
7 (8) Unable to move the computer and/or the other equipment to a comfortable position.
2 (9) Center staff did not respond to my questions or concerns promptly.

12. How long did it take from the time you mailed your registration form to ETS until the time you received your authorization voucher?

| | | |
|---|---|---|
| 20 | (1) | Not applicable (standby) |
| 4 | (2) | Less than a week |
| 41 | (3) | 1 to 2 weeks |
| 26 | (4) | 2 to 3 weeks |
| 5 | (5) | 3 to 4 weeks |
| 1 | (6) | More than 4 weeks |
| 1 | (7) | Did not receive the voucher |

13. Which of these materials would have been helpful to you as you prepared to take the test on a computer? Select as many as apply.

| | | |
|---|---|---|
| 30 | (1) | None - I would not have needed any preparation to take the test. |
| 38 | (2) | Tutorials available on computer |
| 25 | (3) | A printed booklet with all the tutorials and examples from each test section. |
| 19 | (4) | Computer familiarization materials specific to CBT available on computer |
| 25 | (5) | An expanded CBT Supplement with more sample test screens and message screens included in the text. |

**The following questions ask you about the Computer Adaptive Test (CAT), which was administered in Section 7.**

| T | V | Q | A | | |
|---|---|---|---|---|---|
| | | | | 14. | Did you have enough time to answer all of the test questions? |
| 24 | 24 | 25 | 24 | (1) | I answered all of the questions but felt rushed to do so. |
| 53 | 64 | 56 | 36 | (2) | Yes, I completed all of the questions without feeling rushed. |
| 22 | 11 | 19 | 39 | (3) | No, I did not have sufficient time to answer all of the questions. |
| | | | | 15. | Did you READ the material describing the CAT before the administration? |
| 65 | 63 | 69 | 63 | (1) | Yes, I received the materials with my authorization voucher and I read them. |
| 19 | 22 | 16 | 19 | (2) | Yes, the administrator gave me the materials before the test administration and I read them. |
| 5 | 6 | 5 | 3 | (3) | No, I received the materials but did not read them. |
| 10 | 8 | 9 | 13 | (4) | No, I did not receive these materials. |
| | | | | 16. | You were not permitted to "Review" during the last (seventh) section. What was your reaction to this test rule? |
| 57 | 56 | 65 | 49 | (1) | Did not care |
| 35 | 37 | 28 | 43 | (2) | Somewhat frustrating |
| 7 | 7 | 7 | 8 | (3) | Very frustrating |
| | | | | 17. | You were not permitted to omit questions during the last (seventh) section. What was your reaction to this testing rule? |
| 67 | 69 | 72 | 60 | (1) | Did not care |
| 26 | 26 | 23 | 32 | (2) | Somewhat frustrating |
| 6 | 6 | 5 | 8 | (3) | Very frustrating |

18. In the CAT, questions of the same type may not be grouped together. For example, you may have been given an analogy question followed by a sentence completion question and then another analogy question. What was your reaction to this way of presenting the questions?

21 30 16 18 (1) Preferred the CAT presentation
41 47 41 35 (2) Would have preferred to see questions of the same type together.
36 23 42 45 (3) No preference

19. Could you tell that during the CAT you were given questions targeted at your ability level?

31 32 34 27 (1) Yes, all questions seemed challenging but neither too easy nor too hard.
36 33 38 37 (2) Most questions seemed challenging.
10 9 9 14 (3) Many of the questions seemed too hard.
3 4 3 0 (4) Many of the questions seemed too easy.
18 22 14 21 (5) I could not tell that the CAT was a different kind of test.

20. In the directions preceding the CAT questions, you were told the minimum number of questions required to compute your CAT score. Did knowing the minimum number of questions required to compute a CAT score affect how you worked through the CAT test?

23 21 18 30 (1) Yes
75 77 79 67 (2) No

21. Please describe the test taking strategies you used while taking the CAT.

_____

_____

_____

_____

Please comment on any aspect of this computer-administered test.

Please return the completed questionnaire to Educational Testing Service in the attached envelope.

Educational Testing Service
Computer-Based Testing Program
Mail Stop 33-V
Princeton, New Jersey 08541

Table B.1
Questionnaire Percentages by Gender*

| ITEM | TOTAL | FEMALE | MALE | ITEM | TOTAL | FEMALE | MALE |
|---|---|---|---|---|---|---|---|
| 1.1 | 3 | 3 | 3 | 9.1 | 13 | 14 | 11 |
| 1.2 | 19 | 24 | 11 | 9.2 | 61 | 59 | 64 |
| 1.3 | 35 | 37 | 31 | 9.3 | 24 | 25 | 24 |
| 1.4 | 43 | 35 | 53 | 10.1 | 73 | 79 | 65 |
| 2.1 | 38 | 33 | 47 | 10.2 | 16 | 13 | 19 |
| 2.2 | 17 | 18 | 16 | 10.3 | 1 | 1 | 2 |
| 2.3 | 4 | 4 | 5 | 10.4 | 10 | 7 | 14 |
| 2.4 | 38 | 43 | 29 | 11.1 | 63 | 61 | 65 |
| 3.1 | 33 | 38 | 27 | 11.2 | 1 | 1 | 1 |
| 3.2 | 8 | 10 | 5 | 11.3 | 1 | 1 | 2 |
| 4.1 | 1 | 1 | 0 | 11.4 | 9 | 11 | 6 |
| 4.2 | 5 | 4 | 7 | 11.5 | 10 | 13 | 6 |
| 4.3 | 26 | 24 | 28 | 11.6 | 15 | 15 | 15 |
| 4.4 | 43 | 45 | 40 | 11.7 | 4 | 5 | 2 |
| 4.5 | 23 | 23 | 22 | 11.8 | 7 | 8 | 4 |
| 5.1 | 8 | 10 | 6 | 11.9 | 2 | 2 | 2 |
| 5.2 | 33 | 36 | 29 | 12.1 | 20 | 16 | 25 |
| 5.3 | 41 | 40 | 43 | 12.2 | 4 | 3 | 6 |
| 5.4 | 15 | 12 | 19 | 12.3 | 41 | 42 | 40 |
| 6.1 | 14 | 19 | 8 | 12.4 | 26 | 29 | 22 |
| 6.2 | 30 | 33 | 26 | 12.5 | 5 | 6 | 4 |
| 6.3 | 28 | 27 | 29 | 12.6 | 1 | 1 | 1 |
| 6.4 | 25 | 19 | 34 | 12.7 | 1 | 1 | 1 |
| 7.1 | 58 | 59 | 56 | 13.1 | 30 | 28 | 34 |
| 7.2 | 36 | 33 | 38 | 13.2 | 38 | 40 | 35 |
| 7.3 | 1 | 1 | 1 | 13.3 | 25 | 28 | 21 |
| 7.4 | 5 | 5 | 4 | 13.4 | 19 | 23 | 13 |
| 8.1 | 55 | 52 | 58 | 13.5 | 25 | 27 | 23 |
| 8.2 | 25 | 28 | 21 | | | | |
| 8.3 | 19 | 19 | 19 | | | | |

* There were 406 female, 289 male, and a total of 698 respondents (3 did not indicate gender).

Table B.1 (continued)
Questionnaire Percentages by CAT and Gender*

| | All CATs | | | VERBAL CAT | | | QUANTITATIVE CAT | | | ANALYTICAL CAT | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ITEM | Tot | F | M | Tot | F | M | Tot | F | M | Tot | F | M |
| 14.1 | 24 | 23 | 27 | 24 | 25 | 22 | 25 | 22 | 29 | 24 | 21 | 30 |
| 14.2 | 53 | 56 | 49 | 64 | 63 | 67 | 56 | 61 | 50 | 36 | 42 | 28 |
| 14.3 | 22 | 21 | 24 | 11 | 12 | 11 | 19 | 17 | 21 | 39 | 37 | 41 |
| 15.1 | 65 | 66 | 64 | 63 | 66 | 58 | 69 | 68 | 70 | 63 | 63 | 63 |
| 15.2 | 19 | 18 | 21 | 22 | 20 | 26 | 16 | 17 | 16 | 19 | 17 | 22 |
| 15.3 | 5 | 5 | 4 | 6 | 7 | 6 | 5 | 5 | 5 | 3 | 3 | 1 |
| 15.4 | 10 | 11 | 9 | 8 | 8 | 8 | 9 | 10 | 8 | 13 | 15 | 10 |
| 16.1 | 57 | 51 | 65 | 56 | 49 | 65 | 65 | 58 | 75 | 49 | 46 | 52 |
| 16.2 | 35 | 41 | 28 | 37 | 44 | 27 | 28 | 35 | 19 | 43 | 44 | 41 |
| 16.3 | 7 | 7 | 7 | 7 | 6 | 8 | 7 | 7 | 6 | 8 | 9 | 6 |
| 17.1 | 67 | 66 | 70 | 69 | 65 | 75 | 72 | 74 | 70 | 60 | 56 | 64 |
| 17.2 | 26 | 28 | 24 | 26 | 30 | 18 | 23 | 21 | 25 | 32 | 34 | 28 |
| 17.3 | 6 | 6 | 6 | 6 | 4 | 7 | 5 | 5 | 5 | 8 | 9 | 6 |
| 18.1 | 21 | 21 | 20 | 30 | 32 | 26 | 16 | 11 | 23 | 18 | 23 | 9 |
| 18.2 | 41 | 44 | 38 | 47 | 51 | 42 | 41 | 47 | 33 | 35 | 32 | 40 |
| 18.3 | 36 | 33 | 41 | 23 | 16 | 32 | 42 | 42 | 43 | 45 | 43 | 49 |
| 19.1 | 31 | 34 | 27 | 32 | 37 | 25 | 34 | 37 | 30 | 27 | 27 | 25 |
| 19.2 | 36 | 33 | 41 | 33 | 28 | 40 | 38 | 33 | 45 | 37 | 38 | 37 |
| 19.3 | 10 | 9 | 12 | 9 | 8 | 9 | 9 | 10 | 9 | 14 | 10 | 20 |
| 19.4 | 3 | 2 | 3 | 4 | 4 | 3 | 3 | 3 | 4 | 0 | 0 | 0 |
| 19.5 | 18 | 20 | 16 | 22 | 22 | 22 | 14 | 17 | 10 | 21 | 23 | 17 |
| 20.1 | 23 | 23 | 21 | 21 | 24 | 17 | 18 | 17 | 21 | 30 | 32 | 28 |
| 20.2 | 75 | 75 | 75 | 77 | 73 | 82 | 79 | 83 | 74 | 67 | 66 | 69 |
| Number | 698 | 406 | 289 | 235 | 138 | 96 | 264 | 151 | 112 | 199 | 117 | 81 |

*Tot=total group; F=female; M=male